

**ADAPTING THE SCHEDULING OF
ILLUSTRATIONS AND GRAPHS TO LEARNERS
IN CONCEPTUAL PHYSICS TUTORING**

by

Michael Lipschultz

B.S., Pennsylvania State University, 2007

M.S., University of Pittsburgh, 2013

Submitted to the Graduate Faculty of
the Kenneth P. Dietrich School of Arts and Sciences in partial
fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH
KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Michael Lipschultz

It was defended on

December 2nd, 2014

and approved by

Diane Litman, Computer Science Department, University of Pittsburgh

Rebecca Hwa, Computer Science Department, University of Pittsburgh

Jingtao Wang, Computer Science Department, University of Pittsburgh

Vincent Aleven, School of Computer Science, Carnegie Mellon University

Dissertation Director: Diane Litman, Computer Science Department, University of
Pittsburgh

ADAPTING THE SCHEDULING OF ILLUSTRATIONS AND GRAPHS TO LEARNERS IN CONCEPTUAL PHYSICS TUTORING

Michael Lipschultz, PhD

University of Pittsburgh, 2015

This research investigates how to schedule multiple graphical representations in a dialogue-based conceptual physics tutor. Research on multiple graphical representations in tutoring suggests either frequently switching representations or fading from concrete to abstract representations. However, other research communities suggest that the best representation or scheduling can be dependent on various student and tutoring context factors.

This thesis investigates whether these factors are important when considering a schedule of representations. Three major hypotheses are investigated. H1: that the best representational format for physics concepts is related to properties of the student and the tutoring context. H2: that it is possible to build models that predict the best representational format using student and tutoring context information. H3: that picking the representational format based upon student and tutoring context information will produce better learning gains than not considering student and tutoring context information. Additionally, this work addresses the question of whether multiple representations produce greater learning gains than a single representation (H4).

A first experiment was performed to both investigate H1 and to collect data for H2. ANOVAs showed significant interaction effects in learning between low and high pretesters and between high and low spatial reasoning ability subjects, supporting the first hypothesis. Using the data collected and features describing student and tutoring context information, models were learned to predict when to show illustrations or graphs. That these models could be learned, produce meaningful rules, and outperformed a baseline supports H2. A

new modeling algorithm was developed to learn these models by augmenting multiple linear regression to consider certain syntactic constraints.

A third study was run to test H3 and H4 and to extrinsically evaluate the adaptive policy learned. One third of subjects had an adaptive scheduling of representations, one third a fixed alternating scheduling, and one third saw only one representation. In support of H3, subjects with high incoming knowledge sometimes perform better when receiving adaptive scheduling over an alternating scheduling, but there are also counter examples. For H4, it is not supported in general: showing only illustrations is best overall, but in some cases some subjects benefit from multiple representations.

TABLE OF CONTENTS

1.0 INTRODUCTION	1
1.1 CONTRIBUTIONS	4
2.0 RELATED WORK	8
2.1 OUTPUT REPRESENTATIONS IN INTELLIGENT TUTORING SYSTEMS	8
2.1.1 Natural Language Representation	8
2.1.2 Visual Representations	8
2.1.3 Multiple Representations	11
2.2 STUDENT MODELING	13
2.3 STUDENT ADAPTATION	14
3.0 DESIGN OF TUTORING STUDIES	17
3.1 BACKGROUND SURVEY	18
3.2 PHYSICS TEXTBOOK	18
3.3 PRETEST AND POST-TEST	19
3.4 TUTORING SYSTEM ARCHITECTURE	19
3.4.1 Problem Solving	20
3.4.2 Dialogue Architecture	22
3.4.3 Post-Problem Discussion	24
3.5 SATISFACTION SURVEY	25
4.0 FIRST STUDY ON DIFFERENCES REGARDING GRAPHS AND ILLUSTRATIONS FOR LEARNERS	28
4.1 METHODS	29

4.1.1	Participants	29
4.1.2	Material	29
4.1.2.1	Paper Folding Test	29
4.1.2.2	Pretest and Post-test	29
4.1.2.3	Tutoring	30
4.1.3	Procedure	30
4.1.4	Statistical Tests	30
4.2	RESULTS	31
4.2.1	H1: Interaction Between Physics Skill and Condition	33
4.2.2	H2: Interaction Between Spatial Reasoning Skill and Condition	36
4.3	DISCUSSION	39
5.0	MODELING STUDENT BENEFITS FROM ILLUSTRATIONS AND GRAPHS	41
5.1	FEATURES	42
5.2	MODELING	43
5.2.1	Learning Models	45
5.2.2	Statistical Tests	47
5.3	RESULTS	48
5.3.1	Two-Factor Models	50
5.3.2	Three-Factor Models	51
5.3.3	Identifying Non-Simple Adaptive Model	52
5.4	DISCUSSION	55
6.0	EVALUATION OF SCHEDULING REPRESENTATIONS	57
6.1	METHODS	57
6.1.1	Participants	57
6.1.2	Material	58
6.1.2.1	Knowledge Components	58
6.1.2.2	Pretest and Post-test	59
6.1.2.3	Tutoring	60
6.1.3	Procedure	61

6.1.4	Statistical Tests	61
6.2	RESULTS	62
6.2.1	Single Representation versus Multiple Representations	63
6.2.1.1	Learning Gains and Satisfaction	63
6.2.1.2	Interaction Effects	64
6.2.1.3	Learning Curves	69
6.2.2	Alternating Representations vs. Adaptive Scheduling of Representations	71
6.2.2.1	Frequency of Switching	72
6.2.2.2	Learning Gains and Satisfaction	76
6.2.2.3	Interaction Effects	76
6.2.2.4	Learning Curves	80
6.2.3	Modeling Revisited	82
6.3	DISCUSSION	87
6.3.1	Lessons Learned	90
7.0	CONTRIBUTIONS AND FUTURE WORK	93
7.1	H1: BEST REPRESENTATION DEPENDENT ON LEARNER AND TU- TORING CONTEXT	94
7.2	H2: BUILDING MODELS TO PREDICT REPRESENTATION	96
7.3	H3 AND H4: BEST SCHEDULING OF REPRESENTATIONS	98
7.4	GENERALIZING RESULTS	100
	APPENDIX A. BACKGROUND SURVEY	103
	APPENDIX B. TEXTBOOK	105
B.1	INTRODUCTION	105
B.2	DISPLACEMENT	106
B.3	VELOCITY	106
B.4	ACCELERATION	107
B.5	GRAVITATIONAL ACCELERATION	107
B.6	RELATIONSHIPS BETWEEN QUANTITIES	108
B.7	TWO DIMENSIONAL MOTION	109
B.8	RELATIONSHIPS BETWEEN QUANTITIES (2D)	110

APPENDIX C. SATISFACTION SURVEY QUESTIONS	113
APPENDIX D. TEST QUESTIONS FOR THE FIRST STUDY	115
D.1 TEST A	115
D.2 TEST B	124
APPENDIX E. FIRST STUDY PROBLEMS AND REFLECTION QUES-	
TIONS	135
E.1 PROBLEM 1: KANGAROO	135
E.2 PROBLEM 2: ARROW	135
APPENDIX F. SUPPLEMENTAL RESULT ANALYSIS FOR STUDY 1 .	139
F.1 TABLES FOR PRETEST SCORE AND STUDY CONDITION ANOVAS	139
F.2 SPATIAL REASONING AND STUDY CONDITION ANOVAS TABLES .	142
APPENDIX G. TEST QUESTIONS FOR THE SECOND STUDY	145
G.1 TEST A	145
G.2 TEST B	165
APPENDIX H. TUTORING PROBLEMS FOR THE SECOND STUDY .	187
H.1 PROBLEM 1: SUPERTANKER	187
H.2 PROBLEM 2: KANGAROO	188
H.3 PROBLEM 3: ARROW	189
H.4 PROBLEM 4: STONE	191
BIBLIOGRAPHY	196

LIST OF TABLES

4.1	Main effects for the five test score measures ANCOVAs between condition, pretest score, spatial reasoning, and gender on the 22 subjects showing learning gains. The post-test score is the estimated marginal mean of post-test score after controlling for pretest score. The p-values indicate the significance of each factor in the model.	34
4.2	Main effects for the five test score measures ANCOVAs between condition, pretest score, spatial reasoning, and gender for all subjects. The post-test score is the estimated marginal mean of post-test score after controlling for pretest score. The p-values indicate the significance of each factor in the model.	35
4.3	Adjusted post-test score given pretest score for the spatial reasoning main effects for the five test score measures ANCOVAs between condition, pretest score, and spatial reasoning on the 22 subjects showing learning gains. The p-values indicate the significance of the difference between the two levels of spatial reasoning ability.	36
4.4	Adjusted post-test scores for the five test score measure ANCOVAs between pretest score and condition (illustration or graph). The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	37
4.5	Adjusted post-test scores for the five test score measure ANCOVAs between pretest score and condition. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs). All 29 subjects are included.	37

4.6	Adjusted post-test scores for all of the test score measure ANCOVAs that showed significant or trend interactions between spatial reasoning score and condition. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	38
4.7	Adjusted post-test scores for all of the test score ANCOVAs that showed significant or trend interactions between spatial reasoning score and condition. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs). All 29 subjects are included.	39
5.1	Low, median, and high values for all features that were median split.	42
5.2	Models learned are compared according to adjusted r^2 values and their 95% confidence intervals. Italicized rows indicate results significantly better than baseline ($p < 0.05$). Bold indicates the best result.	49
5.3	Rules for the best two-factor model: PreScore.	50
5.4	Rules for the best three-factor model: PreScore*Gender.	52
5.5	Rules for the fifth-best model: PreScore*PctThruSession.	55
6.1	Distribution of students across conditions and sub-conditions.	62
6.2	Comparing single representation to multiple representations across learning gains and satisfaction using independent samples t-test. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5. For all tests, the degrees of freedom are 84.	65
6.3	Comparing constant-graph, constant-illus and multiple representations across the significant measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	65
6.4	Post-hoc comparisons between constant-graph, constant-illus and multiple representations across the significant measures of learning and satisfaction. For each measure, the top value in the cell represents the difference of the average of the two groups. The middle value is the p-value and the bottom value is the power. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	66

6.5	Examining interaction effects between a median split of pretest score and number of representations seen. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	67
6.6	Examining interaction effects between high school physics experience and number of representations seen. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	67
6.7	Examining interaction effects between Calculus experience and number of representations seen. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	68
6.8	Examining interaction effects between gender and number of representations seen. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	69
6.9	Interaction effect results for pretest and single vs. multiple representations on learning curves.	71
6.10	Interaction effect results for high school physics experience and single vs. multiple representations on learning curves.	71
6.11	Interaction effect results for calculus experience and single vs. multiple representations on learning curves.	72
6.12	Interaction effect results for gender and single vs. multiple representations on learning curves.	72
6.13	Number of decision points during tutoring where the representation changed. . . .	73
6.14	Distribution of how often the alternating scheduling picked the same representation that the adaptive policy would have picked.	73
6.15	Correlations between number of times the alternating condition agreed with the adaptive model's prediction and learning or satisfaction.	74
6.16	Correlations between number of times representations switched and learning or satisfaction, on alternating and adaptive conditions.	75
6.17	Correlations between number of times representations switched and learning or satisfaction, on the adaptive condition.	75

6.18 Comparing adaptive scheduling to fixed scheduling of multiple representations across measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5. For all tests, the degrees of freedom are 54.	76
6.19 Interaction effects between a median split of pretest score and scheduling of multiple representations for measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	77
6.20 Examining interaction effects between high school physics experience and scheduling of multiple representations for measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	78
6.21 Examining interaction effects between Calculus experience and scheduling of multiple representations for measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	79
6.22 Examining interaction effects between gender and scheduling of multiple representations for measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.	79
6.23 Interaction effect results for pretest and alternating vs. adaptive scheduling of representations on learning curves.	80
6.24 Interaction effect results for high school physics experience and alternating vs. adaptive scheduling of representations on learning curves.	81
6.25 Interaction effect results for calculus experience and alternating versus adaptive scheduling of representations on learning curves.	81
6.26 Interaction effect results for gender and alternating vs. adaptive scheduling of representations on learning curves.	81
6.27 Low, median, and high values for the new representation features that were median split.	83
6.28 Models learned are compared according to adjusted r^2 values and their 95% confidence intervals.	85
6.29 Rules for the PreScore*Gender baseline.	87
6.30 Summary of findings from the chapter.	92

B1	Equations and descriptions for one dimensional motion.	108
B2	Equations and descriptions for the horizontal component of projectile motion.	111
B3	Equations and descriptions for the vertical component of projectile motion.	112
F1	Raw learning gain pretest-condition ANOVA results for the five test score measures. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	139
F2	Normalized learning gain pretest-condition ANOVA results for the five test score measures. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	140
F3	Raw learning gain pretest-condition ANOVA results for the five test score measures. All 29 subjects are included in this table. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	140
F4	Normalized learning gain pretest-condition ANOVA results for the five test score measures. All 29 subjects are included in this table. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	141
F5	Raw learning gain spatial reasoning-condition ANOVA results for the five test score measures. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	142
F6	Normalized learning gain spatial reasoning-condition ANOVA results for the five test score measures. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	143
F7	Raw learning gain spatial reasoning-condition ANOVA results for the five test score measures. All 29 subjects are included in this table. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	143
F8	Normalized learning gain spatial reasoning-condition ANOVA results for the five test score measures. All 29 subjects are included in this table. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).	144

LIST OF FIGURES

2.1	A cartoonish illustration representing the concept of $1/3$ within the fraction tutor from [92]. Here, one third of the rod has a cylinder around it.	9
2.2	A line graph representing the concept of $1/3$ within the fraction tutor from [92]. Here, one third of the number line is grayed.	9
2.3	Two illustrations depicting the same concept: competitive specialization, from [57].	12
3.1	The practice problem given to all subjects in the studies.	21
3.2	An example of the illustration provided by Andes and the illustrations and graph used in the studies described in this dissertation.	23
3.3	Physics problem solving environment. Andes is on the left and the coach is on the right.	24
3.4	The beginning of the discussion from the first question of the kangaroo problem from Study 2.	27
E1	The two possible visual representations for the kangaroo problem are shown.	137
E2	The two possible visual representations for the kangaroo problem are shown.	138
H1	The illustration presented with the supertanker problem statement.	188
H2	The graph presented with the supertanker problem statement.	189
H3	The illustration presented with the kangaroo problem statement.	190
H4	The graph presented with the kangaroo problem statement.	191
H5	The illustration presented with the arrow problem statement.	192
H6	The graph presented with the arrow problem statement.	193
H7	The illustration presented with the stone problem statement.	194

H8	The graph presented with the stone problem statement.	195
----	---	-----

1.0 INTRODUCTION

American students perform poorly in science compared to their peers in other nations [88]. This has been identified as a problem, with science being considered one of the “core academic subjects” that America needs to focus on and improve student performance on [2]. Despite efforts starting in 2002 to improve education, no significant improvement in American science scores between 1995 and 2007 has been observed. There is still a need to improve science learning in America [1].

One difficult area within science education is physics. Students starting their physics education begin with preconceptions, often misconceptions, regarding physics concepts [103]. During their physics education, students often rely on their “gut feeling” to answer conceptual physics questions [100]. At the end of their physics education, students will often still have those misconceptions [60].

One method used to improve science learning is one-on-one tutoring with a human tutor. This form of tutoring is one of the most effective methods of instruction [30, 71], found to raise student grades two standard deviations above classroom instruction alone [21]. However, human tutors can be expensive and may not be available at the times or for as long as students need. Researchers are developing intelligent tutoring systems (ITSs) to provide similar one-on-one tutoring, but from a computer tutor, which could be always available to the learner. Tutoring systems provide instructional assistance to a learner [105]. ITSs have shown to be effective at improving student knowledge [109, 10, 68], but not as effective as human tutoring.

Attempts to improve physics education through the use of intelligent tutoring systems have targeted both quantitative knowledge (e.g. how to solve physics problems) and conceptual knowledge (e.g. the concept that acceleration is the rate of change of velocity). So

far, developing ITSs for conceptual physics has been found to be difficult. Researchers have developed physics tutoring systems that aid students in learning how to solve quantitative physics problems [109]. However, learning how to solve quantitative physics problems does not necessarily lead to a good understanding of physics concepts [60]. Failing to learn physics concepts correctly can lead to difficulties in learning physics. Students solving a problem correctly, but using incorrect conceptual knowledge, have been shown to struggle with future problems, both problems similar to the original and problems not directly relying on those concepts [99].

To address conceptual physics learning, intelligent tutoring system researchers have taken different approaches to tutoring physics concepts. One dimension to consider is the format used to represent the concepts. Some use a natural language representation, whether through dialogues [108, 78, 107] or canned text [95], to present the concepts. Others use visual representations, either graphs or illustrations, to convey the concepts. Sometimes they are presented with text [7, 58] while other times the visuals are provided with almost no text [62, 3, 4]. Still others have found switching between visual representations during tutoring to be better than using just one visual representation [59, 91]. Section 2.1 introduces representations and discusses how they have been used with conceptual physics tutoring systems.

An orthogonal consideration is how adaptive to make the tutoring system. Student adaptation occurs when a tutoring system’s behavior changes based on the action or actions of the student. Often, the tutoring system consults a student model to identify the student’s behavior or to decide how to change. A student model represents information about the student and can be used to predict what the student knows or how they might behave [105]. The tutoring system can then consult an adaptive policy to determine how its behavior should change given information about the student and the tutoring context. Research has found that adapting to students and tutoring contexts can be more effective than non-adaptive systems (discussed in Section 2.2) and adapting to students is seen as one of the grand challenges of education technology [117]. However, adapting visual representations to individuals is still largely unexplored. In this work, I explore augmenting a natural language-based tutoring system for conceptual physics by adding illustrations and graphs to

the dialogue. In the final study, the tutor adapts to the students and tutoring context by selecting to show illustrations or graphs, depending on which is expected to produce greater learning gains.

The considerations of representational formats to use and how adaptive to make the system lead to the overarching hypotheses of this work: 1) that the best representational format for physics concepts is related to properties of the student and the tutoring context, 2) that it is possible to build models that predict the best representational format using student and tutoring context information, and 3) that picking the representational format based upon student and tutoring context information will produce better learning gains than not considering student and tutoring context information. In testing the third hypothesis, we also investigate a fourth: 4) that showing multiple representations will produce better learning gains than only showing a single representation. These hypotheses were tested through an initial study, where students solved problems in the Andes physics problem solving coach [109] then discussed concepts related to each problem in the Cordillera tutoring system [107]. In one condition, students saw only graphs and in the other, students saw only illustrations. Learning gains were measured using pretests and post-tests. Student information was collected from a background survey and a spatial reasoning test was used to measure their spatial reasoning ability. ANOVAs showed significant interaction effects in learning between low and high pretesters and between subjects with high and low spatial reasoning ability, supporting the first overarching hypothesis. Using this data, models were developed to predict when illustrations and when graphs would most improve learning. That these models could be learned and produce meaningful models that outperformed a non-adaptive model supports the second overarching hypothesis. To create these models, I developed a new modeling algorithm by augmenting multiple linear regression to consider certain syntactic constraints. One of these models was then used in another study, where it was incorporated into the Andes and Cordillera systems. This study compared three conditions: one where subjects saw the same representational format throughout tutoring, one where the representational format alternated according to a pre-set pattern, and one that consulted the model to determine which representational format to use. I find that in some cases adapting to student and tutoring context information produces better learning and

satisfaction. However, there are also instances when that is not the case. I suspect that this may be due to limitations in the model and preliminary investigations provide some support for this.

The following chapters expand on the process and findings from the studies summarized above. Chapter 2 provides a background on the representations used in intelligent tutoring systems and student modeling and adaptation. Chapter 3 describes the general experimental setup for both of the tutoring studies presented in this work. In Chapter 4, a first study was conducted to determine whether there are differences regarding whether graphs or illustrations lead to greater learning gains. Chapter 5 uses the data collected in the first study to model the best representation for a given student and tutoring context. In Chapter 6, one of those models is used in an extrinsic study to determine whether having an adaptive policy for showing illustrations and graphs during tutoring leads to improved learning. Chapter 7 discusses contributions to the fields of Artificial Intelligence and Education, Intelligent Tutoring Systems, and Educational Data Mining. Finally, this chapter considers future work that may follow from this research.

1.1 CONTRIBUTIONS

With different representations providing different benefits to students [81, 69, 84] and there being individual differences between students for representational understanding [84, 70, 39], there appears to be a need for tutoring systems to adapt representations to students. However, tutoring systems that adapt to students do not use multiple representations as part of the adaptation, and tutoring systems that use multiple representations do not adapt those representations and their presentation to individual students.

In this research, I augment an existing natural language-based intelligent tutoring system in the physics domain by adding visual representations – graphs or illustrations – that are adapted to students. This system advances current tutoring systems by incorporating multiple representations – natural language, graphs, and illustrations – and adapting to differences among students such that the visual representation thought to most-improve learning within

a particular tutoring context is displayed. I argue that adapting visual representations to students and tutoring contexts will show greater learning gains for students over a tutoring system that does not adapt. The adaptation is facilitated by a student model, which uses features identified as important in the visual adaptation literature [84, 70, 39], student modeling and adaptation literature [44, 115, 51, 18, 87, 45], and physics education literature [84, 104] discussed above. Learning such a model, when the best representation is not known from the training data, is done through a novel algorithm which extends stepwise regression by considering syntactic constraints and feature relationships. This research benefits the artificial intelligence in education community by identifying features useful for modeling students and predicting whether graphs or cartoonish illustrations are more beneficial to students within the context of a natural language-based intelligent tutoring system for conceptual kinematics.

Chapter 4 shows that there are differences among students regarding whether graphs or illustrations lead to greater learning gains. In a small study, I divide subjects into two conditions, those who will see illustrations during tutoring and those who will see graphs. Tutoring consisted of solving a physics problem in the Andes intelligent homework helper [109], then discussing concepts related to the problem with a natural language-based tutoring system; this was repeated for a second problem. Using learning gains as the dependent variable, I investigate interactions between incoming physics skill, spatial reasoning ability, and visual representation shown. I find that there are interactions, suggesting that there are individual differences among students regarding illustrations and graphs. Prior to this study, it was not known within the artificial intelligence in education community whether such differences existed. This also contributes to the information visualization community by providing further evidence that there are individual differences regarding the best visualization for a task.

In Chapter 5, data collected from the study was used to develop a student model for predicting which visual representation is beneficial for learning. During the study, information was collected that has been found to be useful in other student adaptation tasks or for identifying what visual representation is beneficial for a subject. These include student demographic features [84, 51, 25, 27, 11, 70], incoming student proficiency and student performance [70, 25, 27, 11], and contextual information during tutoring [44, 51, 25, 27, 11].

Using this data, tutoring contexts were identified when graphs were predicted to improve learning gains and tutoring contexts when illustrations were predicted to improve learning gains. These tutoring contexts can then be used to create an adaptive policy. An intrinsic evaluation of the model is performed in this chapter; an extrinsic evaluation of the model is performed in the next. This chapter also presents a technique for mining data to create an adaptive policy when a gold standard is not available. It starts with a standard method (stepwise linear regression) and augments it to remove unwanted terms. The method seeks to identify tutoring contexts when one graphic is better than the other.

Finally, with a student model, a second study was run to test whether adapting to students and tutoring contexts shows greater learning gains than a non-adaptive tutoring system (see Chapter 6). This study uses the same tutoring set up as the first experiment (problem-solving in Andes with post-problem conceptual discussions). In this new study, there were three conditions. One condition, the adaptive condition, uses the student model to determine whether to use graphs or illustrations at points during tutoring. The second condition, the alternating condition, alternates between graphs and illustrations throughout tutoring. Students in the third condition, the constant condition, use only one visual representation throughout tutoring, to be selected randomly at the start of the student's session. Here, the goal is to identify whether adapting to students provides greater learning gains over the other two conditions. Comparing the adaptive and alternating conditions to the constant condition shows whether multiple visual representations is better than a single visual representation in a natural language-based physics tutoring system. This expands on the current research regarding multiple representation during tutoring to include varying visual representations within a natural language-based tutoring system. Results suggest that multiple representations are sometimes helpful, but in other cases a single representations – illustrations – is better. Since the subjects in the study were novices, this finding is consistent with the literature, which suggests that novices benefit most from illustrations [81, 83, 57]. Comparing the adaptive condition to the alternating condition will show whether adapting to student and tutoring contextual differences when determining what visual representation to use is beneficial over just alternating between the two. This is intended to answer the question of whether adaptation is necessary when deciding what visual representation to

use. It will also provide an extrinsic evaluation of the student model. Results suggest that in some cases, adapting is beneficial but in other cases there is no improvement. The lack of strong support for adapting may be a failing of the model which was learned on limited data, evidence that adapting is not beneficial, or that frequency of switching is important to consider (as suggested in [89, 91]). A brief exploration into improving the model with additional features shows that the model had shortcomings. Finally, this study contributes to the information visualization community because there is evidence that adapting visualizations to users improves task performance.

2.0 RELATED WORK

2.1 OUTPUT REPRESENTATIONS IN INTELLIGENT TUTORING SYSTEMS

Researchers have used various representations in tutoring systems to convey information to students. Representations can take the form of natural language or visuals, such as illustrations and graphs. Below are descriptions and examples of each.

2.1.1 Natural Language Representation

Natural language-based tutors use words to communicate ideas between tutor and student, either through dialogues [108, 107, 59, 78, 61, 48, 9, 37, 94] or canned text [95, 34]. Some tutors augment quantitative problem-solving environments to provide conceptual instruction as part of the problem solving (e.g. [95]), while others attempt to teach the concepts directly [108, 78, 107, 59, 61].

While these tutors have shown promise in improving students' conceptual understanding, they do so by relying on words to describe situations that are visual. However, standard physics instructional material (e.g. [120]) rely on visuals (i.e. graphs and illustrations) to teach concepts and even physics experts [103] use visuals to understand physics situations.

2.1.2 Visual Representations

Tutoring systems that use visual representations attempt to convey ideas through the use of visuals. I will focus specifically on two kinds of visual representations commonly used in the tutoring system community: illustrations [59, 92, 101, 56, 73, 96] and graphs [74, 92].

I selected these both because they are common within the tutoring system community and because they are common in physics education [120, 55]. The illustration representational format uses images, whether realistic renderings [58] or cartoonish images [92, 70, 57], to represent situations and concepts. This representation method is intended to present something students can relate to in their everyday lives. The graph representation conveys situations and concepts primarily through graphs, such as bar graphs or line graphs [74, 92]. An example of a cartoonish illustration and a line graph can be seen in Figures 2.1 and 2.2 respectively.

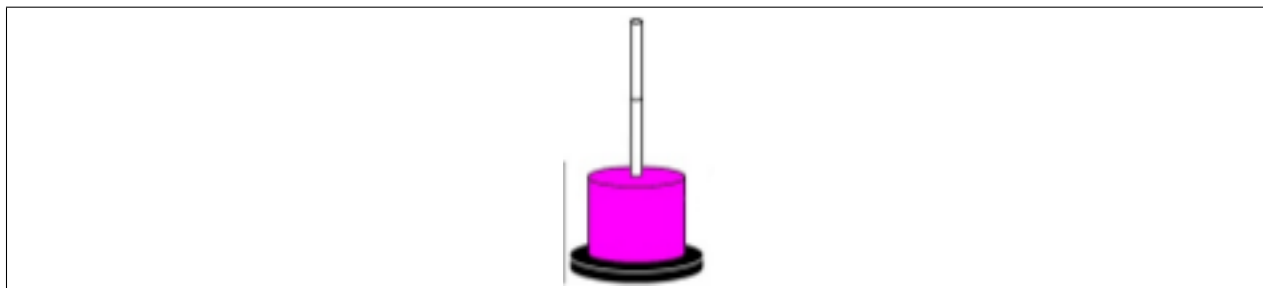


Figure 2.1: A cartoonish illustration representing the concept of $1/3$ within the fraction tutor from [92]. Here, one third of the rod has a cylinder around it.

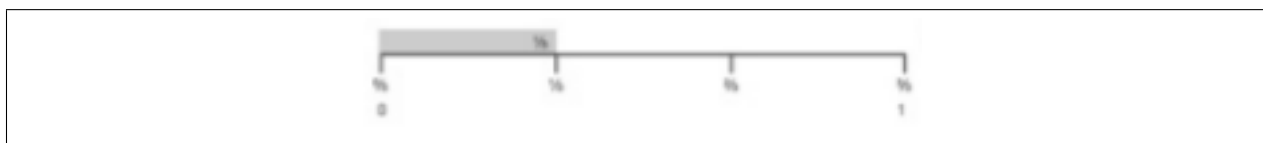


Figure 2.2: A line graph representing the concept of $1/3$ within the fraction tutor from [92]. Here, one third of the number line is grayed.

While there are many different kinds of illustrations, this work specifically focuses on cartoonish illustrations. This form is used in Andes [109], the problem-solving homework helper I will be using in the two studies described later. Additionally, cartoonish illustrations are often used in physics textbooks to convey concepts [120]. The PHeT interactive physics simulations have found that cartoonish illustrations are an effective method of representing physics situations [3].

Tutoring systems using graphs or illustrations provide to the student a visual representation of the situation, sometimes annotated with vectors to further illustrate a concept. The level of interactivity with the visuals varies from system to system, from static images

[7] to interactive simulations that allow the student to manipulate variables in the situation [58, 3, 4]. The optimal level of interactivity with visuals in a tutoring system is not clear. For this work, I selected a low level of interactivity to simplify the interactions between the learner and the system. Allowing for more interactivity could increase cognitive load in the learners and hamper learning. Findings in Chapter 4 suggest this may have been happening with some learners with even low interactivity. It would be interesting in future work to determine how interactive to make visuals when using multiple graphical representations.

The visuals may occur with substantial text accompanying them, such as to provide conceptual help during problem solving [7] or as part of a dialogue with a computer-based tutor [58]. They may also occur where there is little more text than a question and labels, with students expected to take the initiative in exploring the simulation and developing their conceptual understanding [62, 3, 4]. In this work they are presented along with tutorial dialogues since dialogues have been shown to be an effective tutoring strategy [66, 58].

Many of these systems – particularly the interactive simulations – expect students to take the initiative in developing their conceptual understanding by explaining the concepts to themselves (that is, self-explaining). While self-explanations are beneficial for learning [8, 29], students will often not engage in self-explanations on their own. Failing to self-explain can be caused by overestimating their understanding of examples or realizing they have a comprehension gap but lack the knowledge to fill it [93]. To ensure students generate self-explanations, they should be prompted [29, 32]. For the work presented in this dissertation, students are prompted occasionally to self-explain as part of the dialogues.

While including visuals is thought to help students learn concepts, much of the intelligent tutoring system research has made the assumption that one visualization is best for all students. However, differences exist between representations. Illustrations are easier than graphs for novices to interpret [81], but students may be distracted by the surface features of illustrations, which graphs lack [69]. There is also evidence showing the benefits of natural language representations, despite a lack of visuals. Situations presented through natural language representations can lead to more correct answers from students and greater confidence in answers [84].

2.1.3 Multiple Representations

With advantages and disadvantages to the representational formats, some researchers have examined the potential benefits of using multiple representations. Being fluent in multiple representations and being able to translate between them are beneficial for students [92, 5]. Graphs, for example, can help students connect illustrations or verbal descriptions of situations to the base concepts behind them [104].

Research into how to present multiple visual representations has found that using a different visual representation for each question or situation is better than using one visual representation for a while before switching [91]. Additionally, students must be prompted to self-explain for learning gains to be greater than when seeing only one visualization [92]. It is believed that the use of multiple visual representations changing frequently encourages students to repeatedly reactivate knowledge about specific representations. While students do not usually make connections between the representations, being prompted to do so improves learning over these connections being made explicit by the tutoring system [90].

When natural language representations are presented with visual representations, both representations tend to be presented synchronously. The natural language representation may be expository, with the visual representation illustrating the concepts being explained [7]. The representations can also be more interactive, with the natural language representation occurring in the form of a dialogue between tutor and student accompanying a static image [59] or an interactive simulation [58]. Just as with multiple visual representations, students should receive guidance in using the system in a way conducive to learning, such as through prompts to self-explain [63].

While thus far I have talked about using multiple different representations during tutoring, researchers have examined using different forms of one representation with differing levels of concreteness. A concrete illustration is an illustration depicting visual and recognizable objects in a specific situation [96]. An abstract, or idealized, illustration is an illustration depicting concepts apart from specific objects [96]. An example of a situation depicted with both a concrete illustration and abstract illustration can be seen in Figure 2.3. In the figure, the concept being depicted is competitive specialization, the idea that parts

of a system can organize themselves without the help of a leader or a centralized plan. The concrete representation shows how ants will each select their own food source – the closest ant-free food source. The abstract representation shows black dots each selecting a unique gray blob. Researchers have found that fading from concrete illustrations to abstract illustrations shows greater learning gains over just concrete illustrations, just abstract illustrations, or switching from abstract to concrete illustrations [57]. Similar results have been found in the arithmetic domain using concrete and abstract mathematical representations [83].

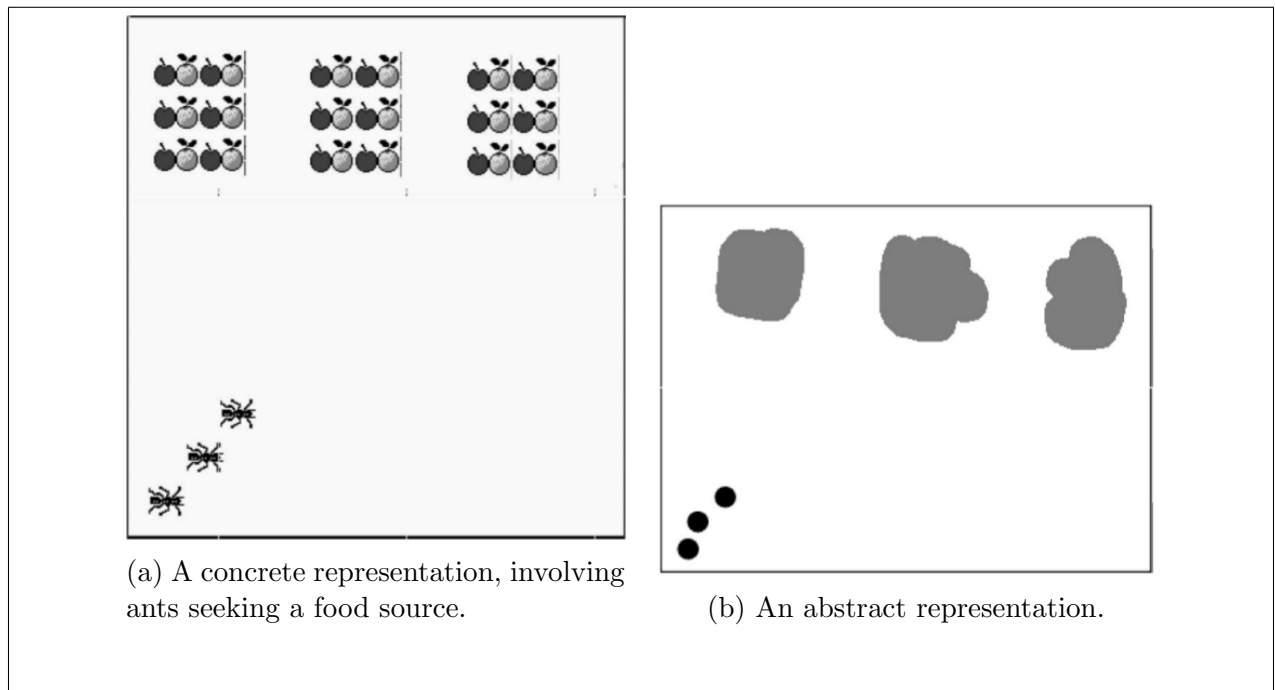


Figure 2.3: Two illustrations depicting the same concept: competitive specialization, from [57].

While these methods of displaying multiple representations during tutoring improve learning, they tend to treat all students as identical. The switching of representations or level of concreteness is determined by a fixed schedule. However, research suggests that there are differences among students that can lead to some visualizations being more beneficial to some students than others, such as gender [84], and spatial reasoning ability [70], verbal skills [39]. Additionally, research suggests that the best visualization can vary for a single student, such as skill with the concepts involved [70], which can change over time, and tasks the student is working on [69]. While tutors using multiple visual representations do not adapt to individual differences, there are other tutoring systems which do adapt to students.

In the next section, I outline student modeling, which is often used in automatically-adapting systems. In the section after, I discuss tutoring systems which adapt to students.

2.2 STUDENT MODELING

Adapting to students can be done with the use of a student model. In this section, student modeling is introduced. Student models are used to infer hidden information about a student, such as what domain knowledge (commonly called “knowledge components” or KCs [105]) a student has mastered or what learning behavior the student is employing. Models infer the hidden information from information readily available to the tutoring system, such as the student’s correctness on previous interactions with the tutor and what KCs the student should know so they can respond correctly [77]. To make these inferences, student models often rely on machine learning or statistical methods [13]. These models are trained on data obtained in a setting and system similar to the ones the model will be used in. The output of these models can then be used to guide tutor behavior, such as whether to provide additional problems to the student [14] or how to respond to a student [6, 112, 77].

One student state often modeled by tutoring systems is student knowledge [35, 38, 22, 86, 24]. Models may take different forms, such as Bayesian networks as in the case of Bayesian Knowledge Tracing [35, 38] or regression models as in the case of Learning Factors Analysis [22, 86, 24]. These models map observables to a prediction of the student’s knowledge. Observables can be such things as information collected from a background survey (e.g. previous experience with the domain) [36], interactions with the tutoring system (e.g. the correctness of previous interactions with the tutor) [86, 18, 87, 45, 51, 51], body language [44, 115], and domain knowledge provided by system designers [22, 86, 24, 35, 38].

Student models have various parameters which must be set. To set them, real-world data is collected containing both the observables and, if possible, the ground truth values that the model will be predicting. In some cases, the ground truth can be observed or directly inferred from an observable, as in the case of whether a student knows a concept, which can be inferred from whether they answer the next question correctly that involves that concept.

In these cases, supervised machine learning can be used to set the parameters. With the model and its learned parameters, it can be intrinsically evaluated on a set of data not used in training the model by measuring how well the model’s predictions match the ground truth values in that held-out set [15, 111]. Here, the ground truth in the data is used as a gold standard to evaluate the model. The model’s performance is compared to a baseline model. In some cases, the ground truth value may not be observable, as in the case of whether it is better to elicit information from a student or to tell them the information [28]. In these cases, other techniques must be used to learn a model and evaluate it, such as optimizing for overall learning gains [25]. In contrast to an intrinsic evaluation, an extrinsic evaluation places the model in a tutoring system and evaluates how well it performs at improving the tutoring. In the elicit versus tell example, an extrinsic evaluation would show that deciding whether to elicit or tell using the learned model improves student learning over some other option (e.g. always telling) [26].

While this section described student modeling in the context of modeling student knowledge, models can also be developed for other student attributes [6, 50, 52, 33], as seen in the next section. In the next section, I discuss in what ways tutoring systems have adapted to learners.

2.3 STUDENT ADAPTATION

Researchers have examined different methods of adapting computer tutors to learners. While some made use of a model (e.g. [33, 44]), others used a Wizard of Oz setup (e.g. [6, 52]). A Wizard of Oz setup replaces a student model with a human, making decisions about the student’s state [6]. This section explores tutoring systems that adapt to learners’ states.

Researchers have had success developing tutors that adapt to students’ mental states, such as motivation [6, 50], uncertainty [52], and various affective states [33]. Adapting to these involves providing additional feedback from the tutor that addresses the student’s disposition (e.g. giving feedback with a positive slant for poorly-performing studious students) [40, 6, 50, 52]. This feedback has led to increased persistence [6], learning gains over no

feedback on affect [52], and increased speed of learning [49]. Detecting these states include such features as body language and facial expressions [44, 115], lexical or dialogue features [44, 51], response times [18], audio or spoken features [87, 45, 51], student features [51], and student goals [31].

Researchers have also examined adapting pedagogical strategies to students. In determining the level of interactivity, it is important to consider the skill level of the student in comparison to the difficulty of the content to learn [106]. For content that is at the current skill level of the student, or easier, high levels of interactivity (e.g. dialogue with tutor) provide no benefit over low or no interactivity (e.g. reading a canned text). For content that is just above the student’s skill level (i.e. material the student has not yet mastered, but is ready to learn), then higher levels of interactivity are beneficial. Determining skill level can be gauged with a pretest and monitored through interactions with the tutoring system [36, 86]. Interactivity can be at the problem-level, where the system will decide which problem to give the student next [36]. Finer-grained level of interactivity decisions (e.g. how intrusive to make hints) can be made using student demographic, temporal, contextual, and student performance features [25, 27, 11].

Students may engage in behavior that is not conducive to learning, such as “gaming the system”, where students obtain correct answers by taking advantage of the tutoring system’s feedback and help [16]. Addressing this behavior has shown to improve learning [14] and has been successfully detecting with models using student performance, contextual, and temporal features [111, 15]. However, not all students who take advantage of the tutoring system’s help are doing so in a manner that hinders learning [98]. Distinguishing between helpful and harmful uses of a tutoring system’s helps can also be done through the use of a model that uses temporal features [98].

Within the realm of adapting representations, some work has been done within natural language-based tutoring systems. When determining how cohesive to make canned text or how cohesive to make tutor turns in relation to student turns in a dialogue, it is important to consider the skill level of the student. Students with low skill levels benefit from high cohesion, whereas students with high skill levels either benefit from low cohesion or there is no difference in the level of cohesion [113, 82, 112].

Similar to the tutoring system community, researchers in information visualization have looked at finding the best representation for the data and for the task, rather than for the user. However, findings suggest that adapting the representation for individual user differences is important. The better representation for a particular user may depend on personality traits, such as locus of control [118] and openness to experience [119]. Visual-spatial abilities, such as visual perception speed [31] and spatial orientation ability [110, 102], can also have an influence on which representation is better. Non-visual cognitive abilities can also have an influence, such as associative memory [23] and domain expertise [41, 102]. With domain expertise, as users gain expertise, the better representation may even change [41]. But even among experts, different visualizations of the same data appear better for experts in different fields [79].

3.0 DESIGN OF TUTORING STUDIES

This chapter introduces the tutoring system architecture and common materials across the two tutoring studies described in this dissertation. The general procedure of both studies is shown below. Both studies were on kinematics, the part of physics involving displacement, velocity, and acceleration. All subjects completed the studies in one session.

1. Complete background survey
2. Read physics textbook
3. Take physics pretest
4. Interface training
5. Tutoring

For each problem in the study:

- a. Solve the problem
 - b. Discuss with tutor each post-problem reflection dialogue
6. Take physics post-test
 7. Take satisfaction survey
 8. Pay subject

In later chapters, “session” will refer to the entire tutoring session, item 5 in the general procedures above. “Problem” will refer to tutoring on a specific problem, both the 5a and 5b in the procedures above. “Problem Solving” specifically refers to 5a and “Post-Problem Discussion” specifically refers to 5b. “WalkThru” refers to the walk-through dialogues the subjects engage in during problem solving.

3.1 BACKGROUND SURVEY

Upon arriving and completing the consent form, subjects completed a background survey, similar to one used in a prior study which also examined post-problem discussions for learning conceptual physics [66]. The change made to the survey was to ask for the three SAT scores; the prior study took place before there were three SAT scores. The survey asked for information thought to be useful for the student modeling task described in Chapter 5, such as gender and whether they had taken high school physics [84, 70]. See Appendix A for the complete survey.

3.2 PHYSICS TEXTBOOK

They then read a short physics textbook on kinematics. The physics textbook was adapted from one used in a previous tutoring study [26]. This previous study was on work and energy, but its textbook included kinematics since kinematics is a prerequisite for work and energy. For the studies discussed in this dissertation, only the sections relevant for kinematics were included. These sections were displacement, velocity, acceleration, and gravitational acceleration. The textbook also used images to illustrate concepts. Since this could act as a confound in the studies presented in this dissertation, the images were removed from the textbook.

Since two-dimensional motion was covered in these studies but were not included in the prior study, a section on two dimensional motion was added. Additionally, kinematics equations were added to the textbook since the subjects in these studies would be solving problems. The complete textbook can be found in Appendix B.

3.3 PRETEST AND POST-TEST

After reading the textbook, subjects took a pretest to measure their incoming physics knowledge. Each study had two isomorphic tests, one randomly selected for the pretest and the other then used for the post-test. The tests were multiple choice to allow for automatic grading. Since each study had its own tests, the details of the tests will be discussed in those chapters.

These tests are administered to measure learning. The tests should be isomorphic to each other; they should ask the same questions but with different cover stories so that the same concepts are evaluated on both tests. The two scores can then be used to measure the learning that took place during tutoring and this can be done in various ways. **Raw learning gain** is the difference between the post-test score and the pretest score; it measures how much improvement there was regardless of how much room for improvement there was. **Normalized learning gain** is the raw learning gain divided by the difference between the maximum score and the pretest score. The numerator represents how much the learner improved; the denominator represents the total room for improvement. Thus, normalized learning gain represents the percent improvement seen. Finally, **post-test score controlled for pretest score** makes use of regression to factor out the effect of pretest score on the post-test score. There is no consensus within the community for which method of measuring learning is best. Some make use of raw learning gains [12, 116], others make use of normalized learning gains [43], others post-test controlled for pretest [54], and still others use multiple measures [53].

3.4 TUTORING SYSTEM ARCHITECTURE

This dissertation made use of the Rimac system[65]. Rimac is a tutoring system for introductory physics. It involves both solving physics problems and a post-problem discussion reflecting on the concepts involved in the problem. Having learners reflect on concepts involved in a problem after solving the problem has been found to be an effective method for

learning physics concepts [66]. Rimac was developed to test the effectiveness of interactivity and levels of abstraction in post-problem reflective dialogues on learning in high school physics students [77]. While the work presented in this dissertation makes use of many of the tools and material developed for Rimac, the research presented here diverges from Rimac on the research questions and target subjects. What follows is a description of the Rimac tutoring system and where my tutoring architecture differs from it.

At the time of running the experiment in Chapter 4, Rimac used the Andes Homework Helper [109] for its problem-solving component. While Rimac later transitioned to having students solve the physics problems for homework, then watch a video showing how to solve the problem before beginning the discussion, this dissertation continues to use Andes for problem-solving. The post-problem discussions were based on the TuTalk Natural Language tutorial dialogue toolkit [64].

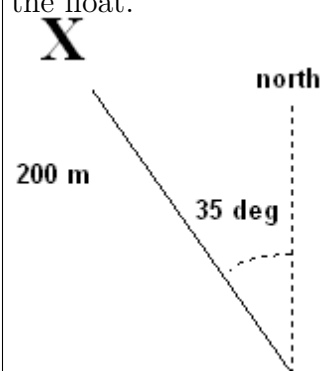
3.4.1 Problem Solving

The Andes Homework Helper is designed to allow students to solve physics problems on a computer. It presents a problem description to the user and an image depicting the situation. Tools are provided to allow the user to draw objects, draw vectors, write equations, and solve for variables. Andes provides immediate color-coded feedback for user actions: green means correct and red means incorrect. At any time, the user may ask for help from Andes and Andes will provide hints on the next step to perform. Occasionally, Andes will offer unsolicited help to the user if the user did something incorrect.

The Rimac project discovered that before using Andes for the first time, both high school physics students and high school physics teachers needed training to become familiar with Andes. Thus, subjects in the studies presented here also underwent interface training. Training began by watching an eight-minute video on how to use Andes, provided by the developers of Andes. It showed the viewer how to solve a simple problem in Andes: the vecIay problem on using vectors. After the video, the user then solved the same problem as in the video to practice using Andes before working on a physics problem. The problem description is shown in Figure 3.1. Solving this practice problem, and all others, was done

with the help of a coach.

A swimmer wishes to reach a swimming float which is 200 m away and 35.0 deg west of north. Find the x and y component of the swimmer's total displacement when she reaches the float.



x component:

y component:

Figure 3.1: The practice problem given to all subjects in the studies.

The physics coach was developed because pilot testing with Rimac found that Andes had a large learning curve. Users continued to struggle with Andes after completing the interface training. The coach would help users both solve the problem and perform actions in Andes. The coach walked users through solving each problem, following the general steps suggested by high school physics teachers involved in the Rimac project:

1. Identify the unknown quantity or quantities the learner must solve for.
2. List the known information that is needed to solve for the unknowns.
3. Identify the necessary equation or equations to solve for the unknowns given the knowns.
4. Solve for the unknowns.

For the practice problem, the coach walked the user through solving the problem, during problem solving the coach would ask the user what they should do instead. If the user was incorrect, the coach would engage in a brief remediation dialogue to guide the user to the correct answer. If the user was correct, the coach would move on. In addition to helping the user go through the problem solving steps, it would also help the user interact with Andes. Whenever the user needed to enter an answer into Andes, the coach would offer help to

remind the user how to perform the action. These problem-solving dialogues were approved by the physics teachers associated with the Rimac project.

Each physics problem in Andes is presented with an illustration of the situation. Since this work compares illustrations and graphs, the default image provided by Andes was replaced with images designed for the studies. Since the graphs developed for the studies represent change over time, the illustrations were updated to also depict change over time. After the first study, it was found that the illustrations did not look enough like illustrations. Thus, for the second study, the Andes illustrations were completely replaced with new illustrations. Figure 3.2 shows the problem description for the kangaroo problem (one of the problems used in both studies described in this dissertation), the original Andes illustration and all of the images created to display in Andes for the studies. All of the illustrations and graphs used in the studies were approved by the physics teachers.

Figure 3.3 shows the problem solving environment. The Andes interface is on the left in the web browser and the coach (running in Cordillera, as discussed below) is on the right. The kangaroo problem from the first study is displayed. The image displayed in Andes is the illustration developed for the study. Cordillera displays the dialogue history between the coach and the user in the top portion while the bottom portion is available for the user to respond.

3.4.2 Dialogue Architecture

The coach, and the reflection dialogues discussed below, were written using the TuTalk toolkit [64]. TuTalk is an authoring tool which enables researchers to write tutoring content through dialogue scripts, which are then used for automated interaction. A TuTalk client runs the script, presenting text and other information (such as images, as discussed below) to the user. It can also elicit responses from the user and take different branches in the dialogue based on the response. The dialogue author lists potential responses and can tag them as either correct or incorrect. The responses can also be tagged for knowledge components, indicating either a user shows evidence of understanding a knowledge component or not understanding a knowledge component.

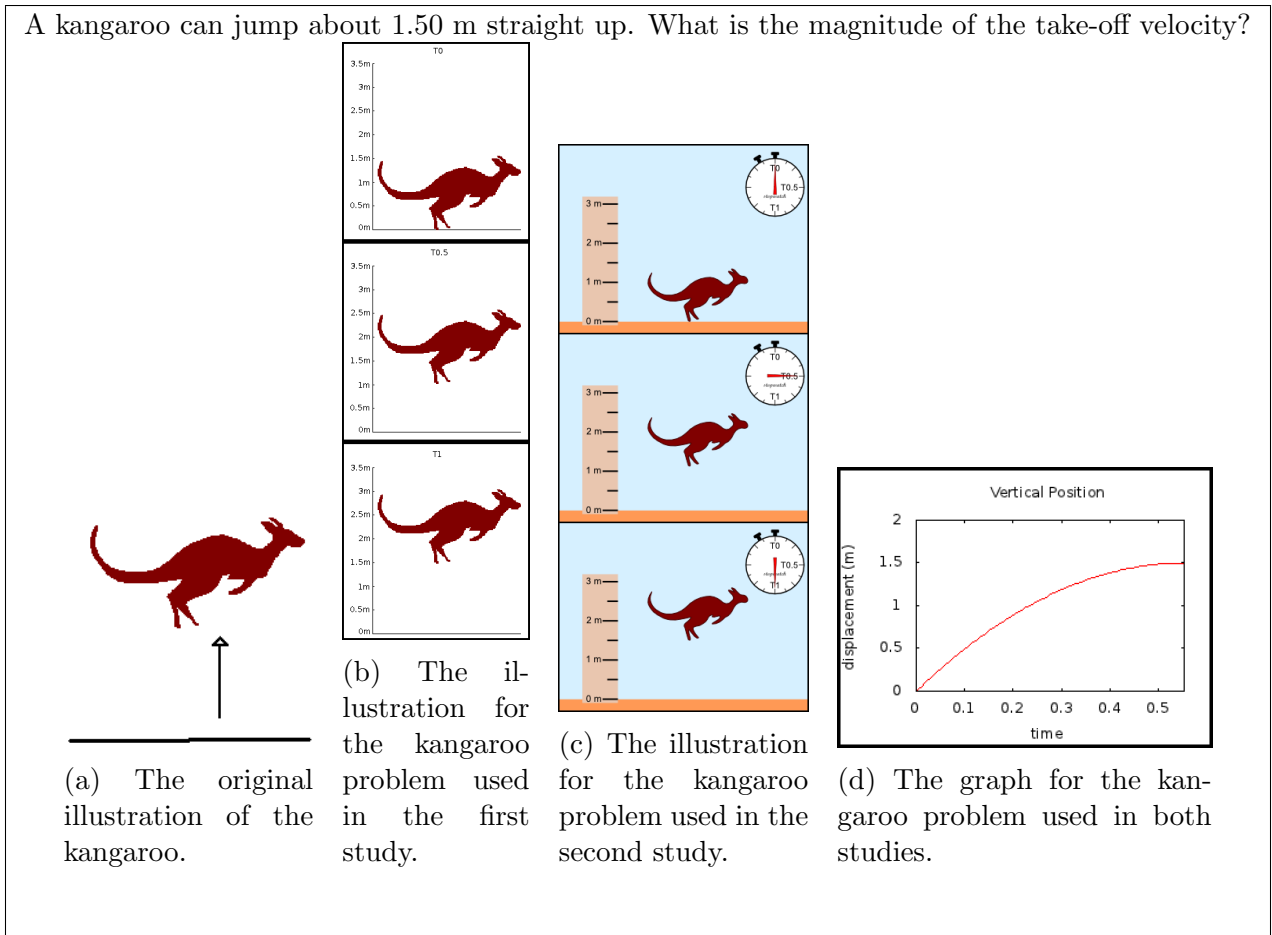


Figure 3.2: An example of the illustration provided by Andes and the illustrations and graph used in the studies described in this dissertation.

The dialogue author can solicit natural language responses from the user or multiple choice responses. For natural language responses, the user's response is compared to a list of anticipated responses and chooses the response that most closely matches the user's response. This classification of natural language responses is prone to misclassifying responses unless the dialogue author lists every possible response from a user. The Rimac project solicited natural language responses since the project was interested in adjusting the tutor's words based on the student's words. However, since my work is interested in scheduling visual representations and not in changing the tutor's word choice, I use only multiple choice questions to eliminate any chance of TuTalk misclassifying a response.

The Rimac project implemented a new web-based interface for the dialogues. However,

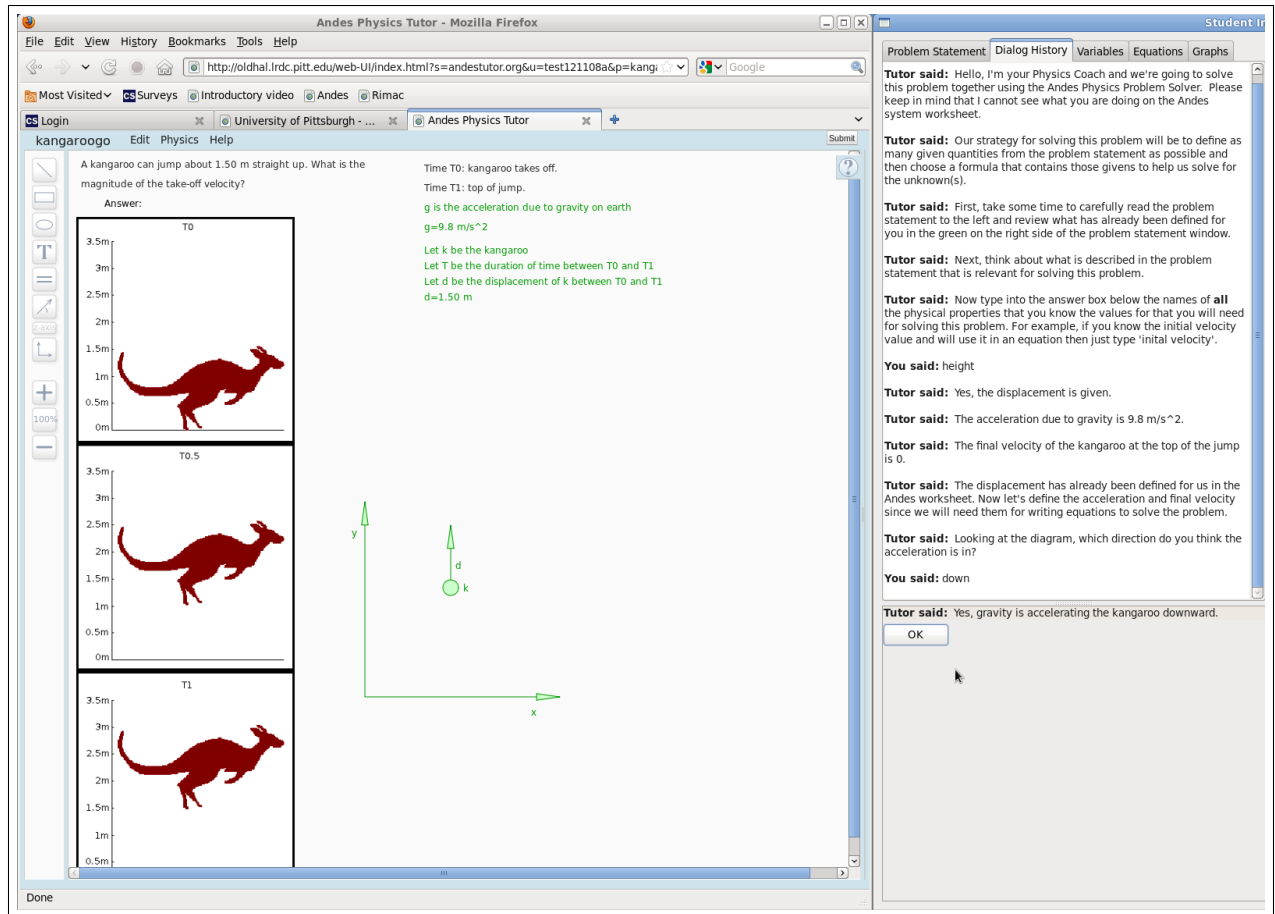


Figure 3.3: Physics problem solving environment. Andes is on the left and the coach is on the right.

for my studies I used the Cordillera tutoring system [107] for both the coach and the post-problem discussion. I chose Cordillera because it had already been tested in a full-scale study [25]. Additionally, it already supported showing graphics alongside the dialogue.

3.4.3 Post-Problem Discussion

Once the user completes a problem, they begin a discussion on concepts involved in the problem. For the first study, described in Chapter 4, these dialogues were written for the study since the Rimac project had not yet settled on the which problems to use. For the second study, described in Chapter 6, the dialogues were adapted from those written for Rimac, which required converting the questions to multiple choice and incorporating graphics

into the dialogues.

The discussion starts by asking the user to reflect upon their solution. For example, the first question asked following the kangaroo problem shown above is “At what time was the kangaroo’s velocity greatest?” After the student answers this question, they begin a natural language-based discussion of the answer. During this discussion, visual representations are shown to the student to help explain concepts. These visuals are shown in a separate pane in Cordillera so that they remain visible as the dialogue progresses. A sample of the discussion from Study 2 for the first question following the kangaroo problem can be seen in Figure 3.4, where graphs are being used.

Each problem has two or three post-problem dialogues associated with it. After the first discussion ends, the next discussion begins. Once all discussions for a problem are completed, the user starts the next problem or moves on to the next part of the study if they have completed all of the problems. All of the post-problem dialogues and their accompanying graphics were approved by the physics teachers involved in the Rimac project.

3.5 SATISFACTION SURVEY

The satisfaction survey was adapted from another study on tutorial dialogue systems [46]. Additional question categories were created specifically for the graphics incorporated into the dialogues. These questions were inspired by those for information visualization [72]. In total there were 27 questions, 23 five-point Likert scale questions and 4 open-ended questions. The Likert scale questions were divided into four categories shown below. All satisfaction questions can be found in Appendix C.

Problem-Solving Graphics – 4 questions on learners’ satisfaction with the graphics during problem solving

Post-Problem Discussion (PPD) – 8 questions on learners’ satisfaction with the post-problem discussions (not including the graphics shown during the discussions)

Post-Problem Discussion Graphics (PPD Graphics) – 6 questions on learners’ satisfaction with the graphics during post-problem discussion

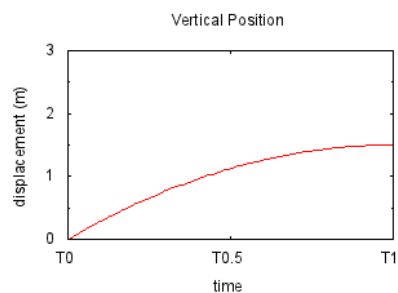
Overall System – 5 questions on learners’ satisfaction with the overall tutoring system

These categorizations allowed us to examine satisfaction with aspects of the tutoring experience. From the four categories, two additional categories were created:

Graphics Satisfaction – a combination of the two graphics categories, totaling 10 questions

Overall Satisfaction – a combination of all four categories, totaling 23 questions

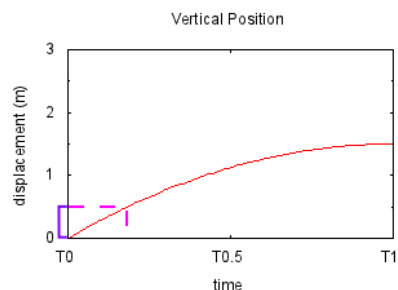
System: The graph on the right shows the vertical displacement of the kangaroo over time. At what time during the jump (from the moment it leaves the ground to the moment it reaches the top of the jump) is the kangaroo moving fastest?



- at T0 (when the kangaroo takes off)
- at T0.5 (half-way through the jump)
- at T1 (top of jump)
- there's no change in velocity
- I am not sure

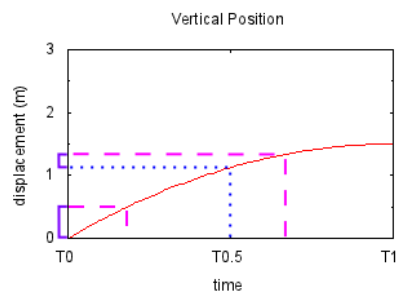
Subject: there's no change in velocity (*incorrect*)

System: Uhm. Velocity is the change in position over the change in time. So let's take a look at the change in position at three instances during the jump: beginning, middle, and end. The beginning of the jump occurs at T0. Since we need two positions to figure out a change in position (that is, velocity), let's also look at a moment (0.1 seconds) after T0.



System: In the figure, a magenta dashed line marks the value of the position at T0+0.1. Notice that the position changes by about 0.5 meters. Do you see the changes in the graph?

- yes
- no



Subject: yes

System: Good. [...]

Figure 3.4: The beginning of the discussion from the first question of the kangaroo problem from Study 2.

4.0 FIRST STUDY ON DIFFERENCES REGARDING GRAPHS AND ILLUSTRATIONS FOR LEARNERS

In this dissertation, I aim to show whether an adaptive policy for showing graphs and illustrations to students during conceptual physics tutoring improves learning. In this chapter, I show that there are individual differences among students regarding the utility of visual representations within the context of a natural language-based conceptual kinematics tutoring system. While others have shown that such differences exist between representations for physics [81, 69, 84], this work has focused on differences when each representation is presented in isolation. It is still unclear whether differences exist between illustrative representations and graph representations when presented along with a natural language representation.

Prior work on using visual representations for learning suggest that differences exist. Therefore, I hypothesize that there is an interaction effect between the visual representation and the student's incoming physics skill. Students who are already skilled in physics should benefit more from graphs, whereas novices should benefit more from illustrations [70, 57]. Skilled students already have some understanding of the physics concepts and so are ready for the more abstract representation of graphs. Novices, however, still need to learn the physics concepts and therefore benefit when the concepts are presented within the context of specific examples, which illustrations provide. Therefore, I hypothesize that novices will show greater learning gains when receiving tutoring with illustrations over graphs (H1a) and that skilled students will show greater learning gains when seeing graphs during tutoring over seeing illustrations (H1b).

There should also be an interaction effect between spatial reasoning skill and visual representation [70]. Students with low spatial reasoning ability should benefit more from illustrations than graphs because it requires greater spatial reasoning ability to imagine what

information about a situation a graph is conveying. High spatial reasoning students should benefit from either visual representation. Therefore, I hypothesize that low spatial reasoning students seeing illustrations will show greater learning gains than low spatial reasoning students seeing graphs (H2).

To determine whether differences exist, I performed an experiment comparing two conditions: one where subjects saw only illustrations during tutoring and the other where subjects saw only graphs. The following section describes the experiment.

4.1 METHODS

4.1.1 Participants

Twenty-nine college students without college-level physics experience were recruited and randomly assigned to one of the conditions. They received \$8/hour plus \$20 for completing the study and having a higher post-test score than pretest score.

4.1.2 Material

Chapter 3 discusses most of the material used in this study. Here, I only discuss the material specific to this study.

4.1.2.1 Paper Folding Test The Paper Folding Test is a standard test for determining spatial reasoning ability [47, 20]. It was administered according to [47] and given following the background survey.

4.1.2.2 Pretest and Post-test There were two isomorphic tests, one randomly selected for the pretest and the other then used for the post-test. Each test had 31 multiple-choice questions. The tests are adapted from two earlier studies, both on post-problem discussions with tutors [66, 65]. Questions were converted to multiple choice and some questions were

modified to include graphics. Four of the questions were problem-solving or numeric questions, all of which were presented using natural language and mathematics. Twenty-seven were conceptual questions. Of the twenty-seven conceptual questions, nine were presented using natural language. Of the remaining 18 conceptual questions, nine involved illustrations and nine involved graphs; these were isomorphic to each other. The complete tests can be found in Appendix D.

By tagging the questions, I was able to measure not just overall learning gains, but also learning gains specifically on problem-solving knowledge (labeled “numeric” in the results), conceptual knowledge, conceptual knowledge in the context of illustrations (referred to as “illustrations” in the results), and conceptual knowledge in the context of graphs (referred to as “graphs” in the results).

4.1.2.3 Tutoring In this study, there were two problems. The first problem involved one dimensional vertical and the second was two dimensional projectile motion. Both problems had three reflection questions. All subjects received the problems and reflection questions in the same order. Appendix E shows both problem statements and their reflection questions.

4.1.3 Procedure

The flow of the study is described in Chapter 3. The Paper Folding Test, which is not listed in the procedures outlined in Chapter 3, occurs between the background survey and the textbook. Before tutoring, subjects were randomly assigned to one of two conditions. In one condition, subjects saw only graphs during tutoring; in the other they saw only illustrations.

4.1.4 Statistical Tests

In analysis of the results below, I make use of various statistical tests. In this section, I discuss the tests.

The Pearson correlation measures the linear relationship between two variables, both the strength of the relationship and whether it is positive or negative. The correlation coefficient, r , ranges from -1 to +1. A coefficient of zero indicates there is no linear relationship between

the variables. A coefficient closer to -1 or +1 indicates a strong linear relationship. A negative value indicate an inverse relationship between the variables, i.e. as one increases the other decreases. A positive value indicates a direct relationship: as one increases the other increases. The coefficient of determination, r^2 , represents how well the statistical model fits the data; higher values are better.

Student's t-test can be one of several tests that follow Student's t distribution. The t-test can be used to determine if two sets of data are significantly different from each other by testing whether their means are significantly different given the standard errors of the two sets of data. In this work, I make use of two t-tests: paired t-test and the independent samples t-test. The paired t-test compares two sets of data, where each item in one set has a corresponding item in the second test. In this work, I use the paired t-test to compare pretest scores to post-test scores, where the pairing occurs with the student, that is the student's pretest score is paired with their post-test score. When it is not possible to pair items from both sets of data, we use the independent samples t-test. In this work, I use it in many cases, such as comparing scores or performance between conditions, where it is not possible to pair a student in one condition with a student in the other. When there is reason to believe the data sets come from different populations (as in the case of some students showing learning gains and others not), Levene's test for the equality of variance should be used to determine whether it is likely the data sets do come from different populations.

The Analysis of Covariance (ANCOVA) tests whether a dependent variable is different across different values of one or more categorical independent variables while controlling for the effects of a continuous variable on the dependent variable. In this work, I use it to identify interaction effects across categorical variables on learning gains after controlling for the subjects' incoming knowledge.

4.2 RESULTS

Twenty-nine subjects participated in the experiment to completion. There were 14 in the illustration condition and 15 in the graph condition.

First, I checked to ensure that the Paper Folding Test measured the kind of spatial reasoning students used in the study. The literature suggests that the test score should correlate with performance on questions involving the interpretation of graphs [70]. However, the Paper Folding Test measures how well students can manipulate a three-dimensional object in their head, which is not similar to the skills needed for interpreting graphs. To test whether the Paper Folding Test is a good measure of graph interpretation, I ran a Pearson correlation test between the paper folding test score and the pretest score on just the nine graph test questions. At a p-value of 0.034, there is a significant correlation between the two scores ($r = 0.3959$). We would also expect that there would not be a correlation with the other test questions. The correlation between the Paper Folding Test and the pretest questions excluding the graph questions has a p-value of 0.0732. Therefore, I conclude that the Paper Folding Test measures the kind of spatial reasoning students used.

I then evaluated whether both conditions were balanced regarding incoming physics experience (as recorded by whether they had high school physics), background physics skill (measured by pretest score), and spatial reasoning ability (measured by the Paper Folding Test). A two-tailed independent samples t-test was run for background physics skill and another was run for spatial reasoning ability. With p-values of 0.943 and 0.524 respectively, I conclude that the two conditions are balanced on both measures. To determine whether incoming physics experience is balanced, I ran a chi-squared test. With a p-value of 0.474, I conclude that the two conditions are balanced on incoming physics experience.

Overall, we see that the subjects learned, with an average pretest score of 18.5 and an average post-test score of 23.2. A paired t-test shows this learning is significant ($p = 0.0006$). Of the 29 students who participated in the study, 22 of them had higher post-test scores than pretest scores. Five subjects in the illustration condition and two in the graphs condition did not show learning gains. In the following analysis, only those 22 students who had learning gains are included. However, for completeness, the results when including all 29 subjects are also shown.

One possible reason those seven subjects did not learn is that they were already high-knowledge subjects. I compared the pretest scores of those who learned to those who did not using an independent samples t-test with equal variances (Levene's test for equality of

variances yields $p = 0.095$). With a p -value of 0.008, we see that there is a difference between the two groups. Those who did not learn had a pretest score of 76.5% and those who did learn had a score of 54.4%. Thus, it appears that some subjects knew too much physics coming in and did not learn more during tutoring.

To test the two sub-hypotheses, ANCOVAs were run to examine interaction effects between condition, pretest score, spatial reasoning, and gender. Five ANCOVAs were run, one for each measure of test score (overall, numeric, conceptual, illustrations, and graphs). Table 4.1 shows the main effects for learning gains across the five test score measures (Table 4.2 shows them for all subjects). For each ANCOVA, the dependent variable was the post-test score, the covariate was the pretest score, and the independent variables were condition (illustration or graph), overall pretest score (median split¹: high or low; median: 58.1%), spatial reasoning ability (median split: high or low; median: 11.5/20), and gender (female or male).

I first examine whether there are any main effects for the three variables of interest, which would suggest an overall effect between levels for that variable. For both condition and pretest, no main effect is seen, suggesting condition and pretest have no overall effect on the differences observed in the post-test measures. We do find main effects for spatial reasoning. Table 4.3 shows the analysis for the spatial reasoning main effects. We see that for all measures of learning, subjects with high spatial reasoning perform significantly (or trend, in the case of the numeric measure of learning) better than those with low spatial reasoning ability.

4.2.1 H1: Interaction Between Physics Skill and Condition

The two sub-hypotheses examined the interaction between physics skill and condition². H1a stated that novices will show greater learning gains when shown illustrations and H1b stated that skilled students will show greater learning gains when seeing graphs. Table 4.4 shows the interaction effects on just the 22 subjects showing learning gains (Table 4.5 shows the interaction effect results for all subjects). Of the five measures of test score, only numeric

¹For all median splits, subjects exactly at the median were categorized as low

²Published in [75]

Table 4.1: Main effects for the five test score measures ANCOVAs between condition, pretest score, spatial reasoning, and gender on the 22 subjects showing learning gains. The post-test score is the estimated marginal mean of post-test score after controlling for pretest score. The p-values indicate the significance of each factor in the model.

	Post-test Measure				
	Overall	Numeric	Conceptual	Illustrations	Graphs
Number of Questions	31	4	27	9	9
Post-test Score	0.768	0.627	0.795	0.793	0.747
Condition	$p = 0.599$ $F_{1,11} = 0.289$ pow. = 0.091	$p = 0.608$ $F_{1,11} = 0.275$ pow. = 0.077	$p = 0.581$ $F_{1,11} = 0.319$ pow. = 0.118	$p = 0.944$ $F_{1,11} = 0.005$ pow. = 0.055	$p = 0.249$ $F_{1,11} = 1.447$ pow. = 0.324
Pretest	$p = 0.321$ $F_{1,11} = 1.059$ pow. = 0.103	$p = 0.196$ $F_{1,11} = 1.844$ pow. = 0.566	$p = 0.418$ $F_{1,11} = 0.696$ pow. = 0.058	$p = 0.260$ $F_{1,11} = 1.378$ pow. = 0.074	$p = 0.399$ $F_{1,11} = 0.757$ pow. = 0.051
Spatial Reasoning	$p = 0.020$ $F_{1,11} = 1.809$ pow. = 0.499	$p = 0.070$ $F_{1,11} = 3.848$ pow. = 0.559	$p = 0.015$ $F_{1,11} = 7.681$ pow. = 0.532	$p = 0.020$ $F_{1,11} = 6.888$ pow. = 0.414	$p = 0.004$ $F_{1,11} = 11.819$ pow. = 0.878
Gender	$p = 0.519$ $F_{1,11} = 0.446$ pow. = 0.093	$p = 0.117$ $F_{1,11} = 2.944$ pow. = 0.342	$p = 0.566$ $F_{1,11} = 0.353$ pow. = 0.084	$p = 0.948$ $F_{1,11} = 0.004$ pow. = 0.050	$p = 0.098$ $F_{1,11} = 3.325$ pow. = 0.378

was neither significant nor a trend.

For sub-hypothesis 1a, regarding low pretesters, we see that this hypothesis is confirmed. Across each of the measures, low pretesters who received illustrations during tutoring scored higher than low pretesters receiving graphs.

Table 4.2: Main effects for the five test score measures ANCOVAs between condition, pretest score, spatial reasoning, and gender for all subjects. The post-test score is the estimated marginal mean of post-test score after controlling for pretest score. The p-values indicate the significance of each factor in the model.

	Post-test Measure				
	Overall	Numeric	Conceptual	Illustrations	Graphs
Number of Questions	31	4	27	9	9
Post-test Score	0.749	0.628	0.772	0.77	0.732
Condition	$p = 0.545$ $F_{1,14} = 0.379$ pow. = 0.069	$p = 0.161$ $F_{1,14} = 2.111$ pow. = 0.117	$p = 0.671$ $F_{1,14} = 0.186$ pow. = 0.067	$p = 0.904$ $F_{1,14} = 0.015$ pow. = 0.051	$p = 0.944$ $F_{1,14} = 0.005$ pow. = 0.110
Pretest	$p = 0.135$ $F_{1,14} = 2.417$ pow. = 0.081	$p = 0.187$ $F_{1,14} = 1.861$ pow. = 0.427	$p = 0.199$ $F_{1,14} = 1.759$ pow. = 0.050	$p = 0.265$ $F_{1,14} = 1.312$ pow. = 0.058	$p = 0.183$ $F_{1,14} = 1.896$ pow. = 0.214
Spatial Reasoning	$p = 0.017$ $F_{1,14} = 6.720$ pow. = 0.395	$p = 0.032$ $F_{1,14} = 5.276$ pow. = 0.446	$p = 0.022$ $F_{1,14} = 6.118$ pow. = 0.352	$p = 0.062$ $F_{1,14} = 3.792$ pow. = 0.180	$p = 0.001$ $F_{1,14} = 14.587$ pow. = 0.706
Gender	$p = 0.419$ $F_{1,14} = 0.695$ pow. = 0.122	$p = 0.200$ $F_{1,14} = 1.807$ pow. = 0.240	$p = 0.522$ $F_{1,14} = 0.432$ pow. = 0.094	$p = 0.662$ $F_{1,14} = 0.199$ pow. = 0.070	$p = 0.583$ $F_{1,14} = 0.316$ pow. = 0.082

Sub-hypothesis 1b, regarding high pretesters, is not generally confirmed. When considering students' overall learning (as shown in Table 4.4), those who saw illustrations during tutoring performed better on overall, conceptual, and graphs questions than those who saw graphs. For all students (see Table 4.5), there is no significant difference for overall and con-

Table 4.3: Adjusted post-test score given pretest score for the spatial reasoning main effects for the five test score measures ANCOVAs between condition, pretest score, and spatial reasoning on the 22 subjects showing learning gains. The p-values indicate the significance of the difference between the two levels of spatial reasoning ability.

Measure	N	Spatial=High	Spatial=Low	p-val	power
Overall	31	0.822	0.685	0.035	0.499
Numeric	4	0.701	0.524	0.099	0.559
Conceptual	27	0.847	0.716	0.028	0.532
Illustrations	9	0.872	0.696	0.033	0.414
Graphs	9	0.814	0.651	0.006	0.878

ceptual questions; for graphs questions seeing graphs during tutoring was better for learning. Since five out of seven subjects who did not show learning were in the illustration condition, this perhaps suggests that the learning for some high pretesters is hindered by illustrations. While the literature suggests that graphs are more helpful for non-novices [70], it does not appear to offer any explanation for why illustrations would be harmful.

When looking at the learning gains on just illustration test questions for the 22 subjects showing learning, we find a trend result that seems to be a reversal. Subjects who saw graphs scored higher than those who saw illustrations during tutoring. Comparing this result to the one from graphs, we see that better performance on graph test questions came from students who saw illustrations during tutoring and better performance on illustration test questions came from students who saw graphs during tutoring. This hints that maybe switching representations improves learning.

4.2.2 H2: Interaction Between Spatial Reasoning Skill and Condition

Hypothesis 2 stated that students with low spatial reasoning ability should show greater learning gains when illustrations are used during tutoring than graphs. Similar to the test-

Table 4.4: Adjusted post-test scores for the five test score measure ANCOVAs between pretest score and condition (illustration or graph). The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	N	Pretest=High		Pretest=Low		Interaction	
		Illus.	Graph	Illus.	Graph	p-val	power
Overall	31	0.876	0.865	0.784	0.655	0.049	0.259
Numeric	4	0.821	0.643	0.791	0.463	0.168	0.544
Conceptual	27	0.876	0.870	0.817	0.694	0.034	0.245
Illustrations	9	0.878	0.916	0.798	0.702	0.080	0.058
Graphs	9	0.828	0.785	0.801	0.634	0.013	0.315

Table 4.5: Adjusted post-test scores for the five test score measure ANCOVAs between pretest score and condition. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs). All 29 subjects are included.

Measure	N	Pretest=High		Pretest=Low		Interaction	
		Illus.	Graph	Illus.	Graph	p-val	power
Overall	31	0.780	0.860	0.747	0.629	0.168	0.105
Numeric	4	0.709	0.732	0.651	0.454	0.329	0.127
Conceptual	27	0.777	0.872	0.782	0.671	0.147	0.100
Illustrations	9	0.793	0.891	0.786	0.694	0.297	0.051
Graphs	9	0.674	0.832	0.759	0.625	0.007	0.211

ing of H1, I again ran ANCOVAs for each of the five test score measures, with independent variables being condition and spatial reasoning test (median split: high, low; median = 11.5/20). Table 4.6 shows the interaction effects for the 22 subjects showing learning gains (Table 4.7 shows the results for all subjects). We find significant results for illustrations

and graphs; the others are trends. Although hypothesis 2 was only interested in low spatial reasoning subjects, I am more generally interested in exploring whether there are spatial reasoning differences among subjects regarding visual representations to address the overarching hypotheses of this dissertation, so I include in this table the high-spatial students as well.

Table 4.6: Adjusted post-test scores for all of the test score measure ANCOVAs that showed significant or trend interactions between spatial reasoning score and condition. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	N	Spatial=High		Spatial=Low		Interaction	
		Illus.	Graph	Illus.	Graph	p-val	power
Overall	31	0.899	0.746	0.658	0.704	0.063	0.244
Numeric	4	0.818	0.584	0.469	0.560	0.073	0.610
Conceptual	27	0.916	0.777	0.697	0.728	0.057	0.110
Illustrations	9	0.926	0.818	0.647	0.728	0.023	0.060
Graphs	9	0.902	0.725	0.662	0.643	0.037	0.420

From these results, H2 is generally not confirmed. Low spatial reasoning subjects appear to perform better on overall, numeric, conceptual, and illustration questions when they are shown graphs during tutoring. This may be due to low spatial reasoning subjects being more challenged by graphs, which require a greater amount of spatial reasoning processing [70]. This challenge may encourage them to try harder to understand the material, thus learning more. The one exception to this is the graphs result for only the 22 subjects showing learning. Here, low spatial reasoning subjects learn better with illustrations. This may be due to graphs being generally difficult for students [19, 84]. Combining the difficulty of learning the concepts while seeing graphs with the difficulty of interpreting the graphs during the post-test, students may have become exhausted or confused, leading to lower scores on these questions.

For high-spatial subjects, the literature suggests that there would not be a difference

Table 4.7: Adjusted post-test scores for all of the test score ANCOVAs that showed significant or trend interactions between spatial reasoning score and condition. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs). All 29 subjects are included.

Measure	N	Spatial=High		Spatial=Low		Interaction	
		Illus.	Graph	Illus.	Graph	p-val	power
Overall	31	0.864	0.732	0.652	0.680	0.229	0.103
Numeric	4	0.815	0.550	0.525	0.544	0.059	0.195
Conceptual	27	0.877	0.769	0.684	0.707	0.351	0.090
Illustrations	9	0.880	0.812	0.697	0.707	0.706	0.050
Graphs	9	0.859	0.725	0.601	0.663	0.066	0.348

between seeing illustrations and seeing graphs [20]. However, we see that those who saw illustrations during tutoring performed better than those who saw graphs. One possibility is that students who saw illustrations during tutoring needed to work harder to learn the concepts in a context-free way. Graphs provide a context-free representation of the concepts and so students in this condition may not have needed to work as hard to learn the concepts in a context-free way. It may also be that since all subjects were novices (since none had taken college-level physics), illustrations were the better graphic to shown based on incoming knowledge, as suggested by some research on the use of multiple graphical representations [57].

4.3 DISCUSSION

This study examined whether there are differences among groups of subjects regarding visual representations within the context of a natural language-based conceptual physics tutoring

system. Showing half of the subjects only illustrations and the other half only graphs, I found that there were interaction effects between learning gains and pretest score and between learning gains and spatial reasoning ability. This suggests that it might be important to adapt visual representations to students.

In particular, it was found that low pretesters benefit from illustrations, which is consistent with the literature [57, 70]. For high pretesters, illustrations during tutoring is helpful for some test questions and no different than graphs for others. However, including the subjects who did not learn, high pretesters score higher on graph test questions when seeing graphs, suggesting that maybe there are instances when graphs are more beneficial than illustrations for high pretesters. Additionally, learning for high pretesters appears to be hindered by illustrations, which is not explained by existing cognitive science or physics education literature.

It was also found that, in contrast to the spatial reasoning literature reviewed earlier, it may be better to show low spatial reasoning students graphs and high spatial reasoning students illustrations. This may be due to low spatial reasoning subjects being more challenged by graphs, which require a greater amount of spatial reasoning processing [70], challenging them to try harder and therefore learn more. However, for graph test questions, low spatial reasoning subjects performed better when being tutored with illustrations. This may be due to graphs being difficult for students [19, 84].

5.0 MODELING STUDENT BENEFITS FROM ILLUSTRATIONS AND GRAPHS

The previous chapter found differences between students regarding illustrations and graphs. An overarching hypothesis this dissertation is exploring is whether adapting to these differences improves learning. Adaptation requires a student model that can predict what kind of visual representation should be shown at various times during a tutoring session. This chapter describes the development of such a model.

Intelligent tutoring systems have used student models to adapt to student knowledge or skill [25, 27, 11, 113], affective states [6, 50, 52], and engagement or interactivity [17, 14]. These models have used such features as domain information [17, 114, 77], student skill within the domain [25, 27, 77], student interaction with the tutoring system and contextual information [14, 44, 51, 18, 87, 45, 51], demographic information [25, 27, 11, 77], and body language and facial expressions [44, 115]. Adaptation can take the form of changes to the presentation of information within a representational format (e.g. highly cohesive text or low cohesion) [113], additional feedback to provide scaffolding [6, 50, 52, 14], or change in pedagogical behavior (e.g. supplementary exercises to cover skipped work) [14, 25, 27]. Both modeling the student and adapting to them have seen success in these other areas, but so far no research has examined adapting visual representations to students.

This chapter explores building models to predict when illustrations and graphs benefit learning. I first describe an algorithm that constructs such models; it uses stepwise linear regression on syntactically-constrained equations, augmented to remove unwanted terms. I then examine the models learned and find that models that include pretest score when describing tutoring contexts perform better. The best model comes from including both pretest score and gender.

5.1 FEATURES

The data for this model learning comes from the study described in the previous chapter. Data from all 29 subject are included, giving 2043 data points at the utterance level, one for each interaction with the coach or reflection tutor. From that data, I extracted the features below. Similar features have been used in previous work on tutoring systems [42, 11, 25] and have been found useful by cognitive science research on visual representations [84, 70]. From this literature, I selected the features I could extract from the data collected during the study.

Median splits were performed for ease of interpreting results. Table 5.1 shows value ranges for the features that were median split.

Table 5.1: Low, median, and high values for all features that were median split.

Feature	Low	Median	High
SpatialReason	1.25	11.50	19.00
PreScore	12.9%	58.1%	100%
WalkThruPctCorrect	0.0%	83%	100%
RQPctCorrect	0.0%	94%	100%
ProblemPctCorrect	0.0%	93%	100%
SessionPctCorrect	0.0%	90%	100%
PctThruProblem	25%	63%	100%
PctThruSession	12%	62%	100%
ItemDifficulty	0.0%	0.0%	100%

Gender – Female or Male

HSPysics – whether the subject had physics in high school or not (**yes**, **no**)

SpatialReason – score on the Paper Folding Test (**high**, **low**)¹

Condition – experimental condition (**graph**, **illustration**)

¹Median splits were performed for ease of interpreting results.

PreScore – score on pretest (**high, low**)

WalkThruPctCorrect – percent of correct answers in the current problem’s walk through dialogue with the physics coach (**high, low**)

RQPctCorrect – percent of correct answers in the current problem’s prior reflection dialogue (**high, low**)

ProblemPctCorrect – percent of correct answers in current problem (both walk through dialogue and prior reflection dialogue(s)) (**high, low**)

SessionPctCorrect – percent of correct answers in tutoring session (**high, low**)

PctThruProblem – for each problem, how far through the dialogues (measured by number of walk through and reflection) the subject has gone (**early, late**)

PctThruSession – how far through tutoring the subject is (measured by number of dialogues completed) (**early, late**)

TimeThruProblem – for each problem, how long (in seconds) the subject has spent with the current problem (both the walk through and reflections) (**early, late**)

TimeThruSession – how long (in seconds) the subject has spent with the tutor (**early, late**)

DialogStage – whether the subject is engaging in **ProblemSolving** or **PostProblemDiscussion**

ResponseTimePrevWords – how long (in seconds) the subject took to respond, divided by the number of words to read in the prompt (**high, low**)

KCusage – whether Knowledge Components (KCs) must be **stated** or **applied**

ItemDifficulty – whether the question is **easy** or **hard**, as determined by percent correct on a small pilot study using these dialogues

5.2 MODELING

I am interested in building an adaptive policy to determine a scheduling of illustrations and graphs that maximizes learning. Standard supervised machine learning technique require a gold standard that indicates the correct answer. In this case, the gold standard would need to

indicate whether an illustration or a graph was the better representation to show the learner at a particular point in tutoring. The data collected from the previous chapter contains the representation a subject saw, but not whether that was the better representation of the two to show them. Since it is not clear how to determine which representation is the better one for a subject at a particular point in tutoring, I cannot use supervised learning to create the adaptive policy.

Instead, I use the subject’s learning gain as an indicator of how beneficial the representation they saw was to learning. To accomplish this, I use an augmented form of stepwise linear regression that seeks to explain the variance in post-test score using interactions between the features above. Standard stepwise regression produces rules that may be contradictory or non-adaptive, which are not helpful in creating an adaptive policy. I augment stepwise regression to address these additional constraints. I also constrain the syntax of the models to better describe the tutoring context. Thus, I am trying to optimize r^2 , subject to certain constraints. This is similar to other regression-based modeling approaches, which place syntactic constraints on the equation [22, 86, 24].

To use the features in stepwise regression, each feature is converted into two factors, one for each value the feature can have, such as *GenderIsFemale* and *GenderIsMale*, for the Gender feature. Each factor has a value of either 1 or 0, depending on whether the feature has that particular value for that data point. For example, for a female subject, *GenderIsFemale* will be 1 and *GenderIsMale* will be 0. From these factors, I develop terms by multiplying two or more factors together, such as *ConditionIsGraph*GenderIsFemale*. Terms are then included in models of the form $postscore = \sum \text{terms} + prescore$. Both *postscore* (post-test score) and *prescore* (pretest score) are continuous variables. *Prescore* is included because pretest scores are often correlated with post-test scores; in this corpus it is a trend ($r = 0.318$, $p = 0.087$).

I use syntactic constraints to identify when illustrations or graphs are better. Others have also constrained the equation syntax when developing regression equations for student modeling [22, 86, 24]. Each term in the model will contain one Condition factor. This is done so that the final model learned can indicate situations when a graphic helped or hindered learning. Additional factors in the term describe the context. For example, if the final model

learned is:

$$\begin{aligned} postscore = & + 0.123 * ConditionIsGraph * GenderIsFemale \\ & - 0.456 * ConditionIsGraph * PctThruSessionIsLate \\ & + prescore \end{aligned}$$

then from this one can deduce the adaptive policy below. A negative sign on the coefficient suggests that it is detrimental to learning to show graphs when it is later in tutoring. Since the only other option is to show illustrations, one can conclude that when it is later in tutoring, show illustrations.

1. When the gender is female, show graphs.
2. When it is later in tutoring, do not show graphs.

One question is what to do for females later in tutoring. In this work, I considered the ordering of the rules and applied the first rule that could be applied. So in the case of females later in tutoring, the “gender is female” rule would be applied since it is ranked higher. However, an alternative would be to add an interaction term to the model and relearn the model. In this case, that would mean adding the terms: $ConditionIsGraph * GenderIsFemale * PctThruSessionIsLate$ and $ConditionIsIllus * GenderIsFemale * PctThruSessionIsLate$. This is one avenue for future work on improving this modeling algorithm.

Because of the simplicity of the example model, it fails to state what to do for males early in tutoring. Learned models were large enough to provide rules for all tutoring situations encountered. Had this not, a simpler model could have been consulted. In the example above, a simpler model could have been one that just used $ConditionIsGraph$ and $ConditionIsIllus$ as terms.

5.2.1 Learning Models

Learning an adaptive policy begins with backward stepwise linear regression. The model learned contains a subset of the terms from the original equation; these terms explain the

variance in post-test score. If one were to extract rules from the model to form the policy, the policy may have contradictory rules. For example, consider the following learned equation:

$$postscore = + 0.123 * ConditionIsGraph * GenderIsFemale \quad (5.1)$$

$$- 0.456 * ConditionIsGraph * PctThruSessionIsLate \quad (5.2)$$

$$+ 0.789 * ConditionIsIllus * GenderIsFemale \quad (5.3)$$

$$+ 0.123 * ConditionIsGraph * PctThruSessionIsEarly \quad (5.4)$$

$$+ 0.111 * prescore \quad (5.5)$$

In this example, the first and third terms contradict each other. The first says to show graphs when the subject is female, while the third says to show illustrations. To address this contradiction, one can remove the term with the lower absolute value of the coefficient² (avc). In this case, remove term 1.

While there are no more contradictions in the equation, I am interested in an adaptive policy and terms 2 and 4 are not adaptive. Together, they say to show graphs regardless of how far along in the tutoring the subject is. As with the contradictory term pair above, remove the term with the lower avc; in this case term 4, leaving the equation:

$$postscore = - 0.456 * ConditionIsGraph * PctThruSessionIsLate$$

$$+ 0.789 * ConditionIsIllus * GenderIsFemale$$

$$+ 0.111 * prescore$$

The final step in learning the equation is to run multiple linear regression to learn the final coefficients of the equation. This step is performed to determine whether any of the coefficient signs change, which could change the conclusions of the policy. In the running example, the model could be:

$$postscore = + 0.012 * ConditionIsGraph * PctThruSessionIsLate$$

$$+ 0.678 * ConditionIsIllus * GenderIsFemale$$

$$+ 0.222 * prescore$$

²I also explored removing both terms, but found that the final models did not perform as well.

To construct a policy from this model, rank the terms according to their avcs and convert them to rules. For example, using the equation above:

1. For female subjects, show illustrations
2. When the subject is later in tutoring, show graphs

The policy can then be applied at every decision point by starting at the top of the list and applying the first that applies. To summarize, the algorithm used to learn the model is:

1. Run stepwise linear regression on the data subject to syntactic constraints
2. Identify problematic term pairs. Problematic rules can be identified by:

Contradictory pair – Two terms concluding opposite conditions and the remaining factors are identical

Non-adaptive pair – Two terms concluding the same condition, the other factors are the same, except one is opposite between the two terms

3. For each problematic term pair, remove the one with the lower avc
4. Using the remaining terms in the model, run multiple linear regression to learn the final model
5. Convert the terms into rules, ranking them from highest to lowest avc

5.2.2 Statistical Tests

In analysis of the results below, I make use of various statistical techniques. In this section, I introduce them.

Cross-validation is a technique for validating a prediction model. Models are often trained on a data set in which both the model inputs and model output are already known. Often, one wants to know how a model will perform on unseen data, to limit overfitting the model to the training data set and to learn how well the model will work on a real-world problem. Cross-validation is a technique for using the training set to both train the model and evaluate it on unseen data. In k -fold cross-validation, the training data is partitioned into k distinct subsets, called folds. Then, the model is trained on $k - 1$ folds and the remaining fold is used for evaluating the model. This training and evaluating is repeated k times, each time with a different fold held out for evaluation. When evaluation is complete for an iteration,

the evaluation metric is recorded and ultimately combined with the evaluation metrics from the other iterations. In this work, k is taken to be 10.

The evaluation metric used to compare the models is the adjusted r^2 value. As discussed in 4.1.4, the r^2 value is a measure of how well a statistical model fits the data (higher values are better). Models with more independent variables may have a higher r^2 but at the cost of increased complexity. The adjusted r^2 considers the number of independent variables when determining the value and will increase if the additional variables improve the predictive ability of the model. Since I compare the models with varying numbers of terms, I compare them on the adjusted r^2 instead of r^2 .

In comparing models using k -fold cross-validation, one has k values for each model. To compare the models, I compute the 95% confidence interval for the true mean of those k values for each model. Confidence intervals for the mean are a range in which one can be $n\%$ confident that the true mean lies within. The confidence interval for one model can be compared to the confidence interval for another model. If there is an overlap, then it can be said that the performance of one model and the performance of the other model may not be different. Thus, confidence intervals can be a method of comparing many models to each other.

5.3 RESULTS

The models are compared to a baseline, which always predicts showing the same kind of graphic³. Since showing illustrations produced slightly better learning gains than graphs, I use as the baseline a model that always shows illustrations. The models are each evaluated using ten-fold cross validation and are compared according to the adjusted r^2 value. The performance of the baseline can be seen in the first row of Table 5.2. While the performance of the models are calculated using 10-fold cross-validation, the model rules presented come from the model learned on all of the data.

³Most of these results have been published in [76]. The 3 Factor results that do not include PreScore are unpublished, as is the work on identifying a non-simple adaptive model (Section 5.3.3).

Table 5.2: Models learned are compared according to adjusted r^2 values and their 95% confidence intervals. Italicized rows indicate results significantly better than baseline ($p < 0.05$). Bold indicates the best result.

Model		Adj. r^2	95% CI
Baseline (Illustration)		0.1127	(0.0896, 0.1358)
1 Factor		0.0955	(0.0737, 0.1172)
2 Factors	<i>Gender</i>	<i>0.1788</i>	<i>(0.1428, 0.2148)</i>
	SpatialReason	0.1488	(0.1149, 0.1826)
	<i>PreScore</i>	<i>0.3499</i>	<i>(0.3266, 0.3732)</i>
	PctThruProblem	0.1007	(0.0635, 0.1378)
	PctThruSession	0.1180	(0.0851, 0.1509)
3 Factors	<i>PreScore & Gender</i>	<i>0.4571</i>	<i>(0.4220, 0.4922)</i>
	<i>PreScore & SpatialReason</i>	<i>0.2817</i>	<i>(0.2367, 0.3267)</i>
	<i>PreScore & PctThruProblem</i>	<i>0.3418</i>	<i>(0.3183, 0.3653)</i>
	<i>PreScore & PctThruSession</i>	<i>0.3087</i>	<i>(0.2782, 0.3392)</i>
	<i>Gender & SpatialReason</i>	<i>0.3500</i>	<i>(0.3193, 0.3807)</i>
	<i>Gender & PctThruProblem</i>	<i>0.2093</i>	<i>(0.1672, 0.2515)</i>
	<i>Gender & PctThruSession</i>	<i>0.1923</i>	<i>(0.1451, 0.2396)</i>
	SpatialReason & PctThruProblem	0.1510	(0.1051, 0.1969)
	<i>SpatialReason & PctThruSession</i>	<i>0.1688</i>	<i>(0.1444, 0.1932)</i>
	PctThruProblem & PctThruSession	0.1068	(0.0677, 0.1459)

The second row in Table 5.2, the “1 Factor” model, presents the model learned when there is only one factor describing the tutoring context, and the interaction feature Condition. That is, the initial equation was of the form shown below, where F_1 is the set of all factors except those involving Condition. This produces a model similar to those seen in the previous section.

$$\begin{aligned}
postscore = prescore \\
+ (ConditionIsGraph + ConditionIsIllus) * \sum_{factor_i \in F_1} factor_i
\end{aligned}$$

The adjusted r^2 value is 0.0955 for this model, which is not significantly different than the baseline. One possible reason for the poor performance of this model is that it can only identify simple tutoring contexts. Since all terms in this model consist of one non-Condition

factor, the model can only identify tutoring contexts by one feature (e.g. *GenderIsFemale* or *PctThruSessionIsLate*). One feature may not be enough to adequately describe tutoring contexts when illustrations or graphs are more beneficial than the other; the descriptions may be too coarse-grained.

5.3.1 Two-Factor Models

To create finer-grained tutoring context descriptions, another factor feature is added to each factor to describe the tutoring context. Five features were selected based on prior work suggesting a change in these features can cause large changes in models [75, 70]: *Gender*, *SpatialReason*, *PreScore*, *PctThruProblem*, and *PctThruSession*. An example of one such initial equation is shown below (for *Gender*), where $F_{2,Gender}$ is the set of all factors except *Condition* or *Gender* factors. To illustrate what the model looks like, one term from the model is *ConditionIsGraph*GenderIsFemale*PctThruSessionIsLater*.

$$\begin{aligned}
 postscore = prescore + & [(GenderIsFemale + GenderIsMale) \\
 & * (ConditionIsGraph + ConditionIsIllus) \\
 & * \sum_{factor_i \in F_{2,Gender}} factor_i]
 \end{aligned}$$

The results for these models can be found in the “2 Factor” models in Table 5.2. Of the five models, two perform significantly better than baseline: *Gender* and *PreScore*, with *PreScore* significantly better than other models seen so far. The rules for the model are in Table 5.3.

Table 5.3: Rules for the best two-factor model: *PreScore*.

High Pretesters (n = 11)	Low Pretesters (n = 18)
1. If WalkThruPctCorrect=Low, show Graphs	1. If SessionPctCorrect=High, show Illus
2. If RQPctCorrect=Low, show Graphs	2. If ProblemPctCorrect=Low, show Illus
3. If SessionPctCorrect=High, show Illus	3. If PctThruSession=Late, show Illus
4. If PctThruSession=Late, show Illus	
5. If ProblemPctCorrect=Low, show Graphs	

From the model, one can see that low pretesters should be shown illustrations. This is consistent with cognitive science research, which suggests that novices benefit from illustrations because they are more concrete and relatable than graphs [70]. It is also consistent with the results from Chapter 4, which found low pretesters benefit from illustrations.

High pretesters, however, benefit when the representations change. This sheds some light on prior work on this corpus, which found no significant interaction effect between Condition and PreScore for high pretesters. While the prior finding suggested the graphic did not matter for high pretesters, the model above suggests tutoring contexts when the graphic does matter. If there are many correct responses in the tutoring session or if it is later in tutoring, illustrations are better. If there are few correct responses during the walk through dialogue, reflection dialogue, or the problem, show graphs.

5.3.2 Three-Factor Models

Given the success of finer-grained tutoring context descriptions by adding a second tutoring context description factor, I now explore adding a third. The third factors were selected from the same set of five as in the previous section.

The results for these models can be found in the “3 Factors” models in Table 5.2. The models perform significantly better than baseline, with PreScore*Gender performing significantly better than the rest. Table 5.4 has the rules for the model.

Within this model, one can see difference between Genders. For example, low pretesting females may benefit from seeing graphs when SessionPctCorrect is high, whereas low pretesting males may benefit from illustrations when SessionPctCorrect is high. Differences also exist between PreScore. For example, female high pretesters may benefit from graphs when PctThruSession is early whereas female low pretesters may benefit from illustrations. That these differences exist in the model suggest that looking at interactions with both features improves tutoring context description.

Comparing this model to 2-factor PreScore model, one can see that there are situations when low pretesters may benefit from graphs. Low pretesting females may benefit from graphs when SessionPctCorrect is high. Low pretesting males may benefit from graphs

Table 5.4: Rules for the best three-factor model: PreScore*Gender.

Female High Pretesters (n = 8)	Female Low Pretesters (n = 9)
<ol style="list-style-type: none"> 1. If WalkThruPctCorrect=Low, show Graphs 2. If RQPctCorrect=Low, show Graphs 3. If SessionPctCorrect=High, show Illus 4. If ProblemPctCorrect=High, show Illus 5. If PctThruProblem=Early, show Graphs 6. If PctThruSession=Early, show Graphs 	<ol style="list-style-type: none"> 1. If SessionPctCorrect=High, show Graphs 2. If PctThruSession=Early, show Illus 3. If ProblemPctCorrect=High, show Illus 4. If PctThruProblem=Early, show Illus 5. If RQPctCorrect=Low, show Illus
Male High Pretesters (n = 3)	Male Low Pretesters (n = 9)
<ol style="list-style-type: none"> 1. If RQPctCorrect=Low, show Illus 2. If SessionPctCorrect=High, show Illus 3. If WalkThruPctCorrect=Low, show Illus 	<ol style="list-style-type: none"> 1. If RQPctCorrect=Low, show Illus 2. If WalkThruPctCorrect=Low, show Illus 3. If SessionPctCorrect=High, show Illus 4. If PctThruSession=Early, show Graphs 5. If PctThruProblem=Early, show Graphs 6. If ProblemPctCorrect=High, show Illus

earlier in the tutoring session or earlier in a new problem. From this, I conclude that increasing the descriptiveness of each term in the model can yield improved performance. Continuing to create finer-grained tutoring context descriptions might yield better models, but the data set is too small to handle the increase in terms that would result.

5.3.3 Identifying Non-Simple Adaptive Model

Since I am interested in showing the effectiveness of adapting representational formats to learners and tutoring contexts, I would like to have a model that is not simple with regard to its adaptive rules. I am seeking a policy that predicts showing illustrations and graphs approximately equally and includes a variety of tutoring contexts.

While parsimony is generally a desirous property of models [97], in the testing of overarching hypothesis 3 of this dissertation (adapting the schedule of illustrations and graphs is better than alternating between them), I prefer a model that has many adaptive rules over a more parsimonious one. For example, a policy with the rules shown below does not capture much variety during tutoring.

1. If PctThruSession=Early, show illustrations
2. If PctThruSession=Late, show graphs

A policy with rules incorporating more features can capture more variety, such as the one shown below.

1. If PctThruSession=Early, show illustrations
2. If Gender=Female, show graphs
3. If SessionPctCorrect=High, show illustrations
4. If PctThruSession=Late, show graphs

More rules does not necessarily indicate more variety. Some features capture similar information. For example, SessionPctCorrect and ProblemPctCorrect measure correctness over a series of dialogues. If one is high, then the other is likely to be high as well, as can be seen for high pretesting females. Thus, the sub-policy for high pretesting females could be simplified by merging the first two rules and the second two rules:

1. If few correct responses, show graphs
2. If many correct responses, show illustrations
3. If PctThruProblem=Early, show graphs
4. If PctThruSession=Early, show graphs

A policy that predicts one representation far more often than the other may result in a fairly non-adaptive policy. In the extreme case, a model that includes many features, but always predicts one representation is no better than a model that includes no features and just predicts that representation. For example, the Female Low Pretester sub-policy in Table 5.4 has five rules but since four of them predict the same representation, it can be simplified to:

1. If SessionPctCorrect=High, show graphs
2. Otherwise: show illustrations

It is also possible to observe trends within a partition, such as the only time to show males graphs is when they are starting the tutoring session or starting a new problem. Thus, the males sub-policy can simplify to:

1. If low pretester and either starting the session or starting a new problem, show graphs
2. Otherwise, show illustrations

Following the above rules simplifies the PreTest*Gender model to:

Female:

Low pretester:

1. If many correct answers during tutoring: show Graphs
2. Otherwise: show Illustrations

High pretester:

1. If many correct answers: show Illustrations
2. Otherwise: show Graphs

Male:

1. If low pretester and either starting the session or starting a new problem: show Graphs
2. Otherwise: show Illustrations

The fifth-best performing model, PreScore*PctThruSession, (shown in Table 5.5) is the best-performing model that has many adaptive rules. For high pretesters, half of the policy rules require knowledge of progress through the tutoring session. For low pretesters all of the rules require that information. For each of the sub-models, many rules are needed to describe the model. Therefore, because of its relative complexity, this is the model used in the next study. Applying the simplification steps outlined above yields:

High pretester:

1. If early in session and few correct answers: show Illustrations
2. If few correct answers: show Graphs
3. If many correct answers: show Illustrations
4. If later in tutoring and early in problem: show Graphs

Low pretester:

Early in session:

1. If few correct answers in problem: show Graphs

Table 5.5: Rules for the fifth-best model: PreScore*PctThruSession.

High Pretesters Early in Session	Low Pretesters Early in Session
1. If RQPctCorrect=Low, show Graph 2. If WalkThruPctCorrect=Low, show Graph 3. If SessionPctCorrect=High, show Illus 4. If ProblemPctCorrect=Low, show Illus	1. If ProblemPctCorrect=Low, show Graph 2. If SessionPctCorrect=Low, show Illus 3. If ProblemPctCorrect=High, show Illus
High Pretesters Later in Session	Low Pretesters Later in Session
1. If SessionPctCorrect=Low, show Graph 2. If ProblemPctCorrect=High, show Illus 3. If WalkThruPctCorrect=Low, show Graph 4. If RQPctCorrect=Low, show Graph 5. If PctThruProblem=Early, show Graph	1. If RQPctCorrect=Low, show Illus 2. If WalkThruPctCorrect=Low, show Illus 3. If SessionPctCorrect=Low, show Graph 4. If SessionPctCorrect=High, show Illus 5. If ProblemPctCorrect=High, show Graph

2. If few correct answers in session: show Illustrations
3. If many correct answers in problem: show Illustrations

Later in session:

1. If few correct answers in dialogue: show Illustrations
2. If few correct answers in session: show Graphs
3. If many correct answers in session: show Illustrations
4. If many correct answers in problem: show Graphs

5.4 DISCUSSION

Prior work on this data found differences from the pooled data [75] by identifying situations when one group of students may benefit from one visual representation over another. This chapter identifies tutoring contexts when one representation might be better than the other and certain partitionings of subjects and tutoring contexts that greatly improve the model's adjusted r^2 value. In particular, partitioning along pretest score or gender produces significantly better models than no partitioning or partitioning along spatial reasoning ability or

percentage through a problem or through the tutoring session. These findings are consistent with the literature, which suggested that gender [84] and domain knowledge [79, 70] are important factors when deciding on a representation to show. That spatial reasoning ability was not an important feature to consider is counter to the cognitive science literature [110, 70].

Examining the rules learned in the best-performing model, as measured by adjusted r^2 value, find instances that appear to agree with the cognitive science literature. Some research suggests beginners should start with illustrations since they are relatable, whereas those with some knowledge might benefit more from graphs [70, 57, 104]. We see this with females in the PreScore*Gender model. Low pretesting females should see illustrations when starting the tutoring session or starting a new problem, unless they have been doing well in tutoring already in which case they should see graphs. High pretesting females should start off with graphs. However, we also find cases that disagree with the literature, such as low pretesting males starting with graphs. For low pretesters, there appears to be an interaction with gender regarding which representation to start with, as suggested by Meltzer’s research which found that females perform worse than males when given abstract representations such as graphs [84].

This chapter also presented a technique for mining data to create an adaptive policy when a gold standard is not available. It starts with a standard method (stepwise linear regression) and augments it to remove terms unwanted for developing adaptive systems. The method seeks to identify tutoring contexts when one graphic is better than the other. Increasing tutoring context descriptions, by adding more factors to each term, improve model performance. Many models, particularly those involving PreScore, significantly outperform the baseline.

Finally, while some models performed significantly better than baseline, they did not learn complex adaptive rules. The fifth best-performing model had the most adaptive policy. This model, PreScore*PctThruSession, is used in the next chapter. In the next chapter, I perform an extrinsic evaluation of the PreScore*PctThruSession by incorporating it into the tutoring system and evaluating its effectiveness at selecting representations that aid learning compared to both alternating visual representations and using only one throughout tutoring.

6.0 EVALUATION OF SCHEDULING REPRESENTATIONS

In the previous chapter, I found tutoring contexts when one representation should be more beneficial for learning than another. From these tutoring contexts, an adaptive policy was developed. In this chapter, I evaluate the PreScore*PctThruSession model by performing a user study. This study will test whether adapting the scheduling of graphs and illustrations to students and tutoring contexts improves learning over an alternating scheduling of graphs and illustrations.

The study is also designed to answer another question within the ITS community. With this study, I wish to show whether multiple visual representations is better than a single visual representation in a natural language-based physics tutoring system. Current research suggests that switching representations regularly improves learning over switching infrequently [91] or not at all [92], but this is within the fractions domain, does not make a distinction between visual representations (e.g. illustrations versus graphs), and is not within the context of a natural language-based tutoring system.

6.1 METHODS

6.1.1 Participants

Eighty-seven subjects without college physics were recruited and randomly assigned to one of the conditions. They received \$8 per hour plus \$5 for completing the study and having a higher post-test score than pretest score.

6.1.2 Material

Chapter 3 discusses most of the material used in this study. Only the material specific to this study is discussed here.

6.1.2.1 Knowledge Components The knowledge components (KCs) used in this study were developed for an earlier study on post-problem tutorial discussions with a human tutor [66], which the Rimac project then extended and used to tag their dialogues [65]. Since the study described in this chapter focuses on a smaller section of physics, the KCs used in this study are a subset of those used in the previous studies. The numbering of the KCs has been kept the same as with those previous studies, thus “missing” KCs are those not used in this study. What follows are the six KCs present in this study, along with a description and example of each. For the dialogue examples, the subject’s response is shown with the correctness tag in italics. Examples of the KCs in the tests can be found in Appendix G.

Definition of Acceleration (KC15) – Acceleration is a change in velocity over a certain time interval. It is a vector, and hence has magnitude and direction and can be defined in terms of `acceleration_x` and `acceleration_y` (for 2-D motion) and its S.I. unit is m/s^2 . An example from the kangaroo problem’s post-problem discussion:

System: What physical quantity represents a change in velocity over time?

Subject: acceleration (*correct*)

Direction of Acceleration and Velocity (KC23) – Acceleration points in the same direction as the change in velocity. If an object is speeding up, acceleration points in the same direction as the velocity; if it is slowing down, it is decelerating and it points opposite the velocity. An example from the kangaroo problem’s post-problem discussion:

System: So we’ve established that the velocity is pointing up, but that it’s getting smaller. Let’s now look at the direction of acceleration. What direction is the acceleration of the kangaroo during that time?

Subject: Down (*correct*)

Gravitational Acceleration (KC24) – The acceleration due to gravity is a downwards-pointing vector with a practically constant magnitude (9.8 m/s^2) (near the earth’s surface)

and therefore only the y-component of acceleration is affected. An example from the kangaroo problem's post-problem discussion:

System: What is the acceleration due to?

Subject: gravity (*correct*)

Horizontal Velocity is Constant in Projectile Motion (KC25) – In ideal projectile motion, the horizontal components of the projectile's velocity are unaffected by gravity, and remain constant. An example from the arrow problem's post-problem discussion:

System: As we discussed earlier, the arrow experienced an acceleration in the y-direction; it was due to gravity. Did the arrow experience gravity in the x-direction?

Subject: No (*correct*)

Definition of Velocity (KC28) – Velocity is displacement (change in position) per time. It is a vector with dimensions of length/time and SI units of meters/second (m/s). An example from the kangaroo problem's post-problem discussion:

System: At what time during the jump (from the moment it leaves the ground to the moment it reaches the top of the jump) is the kangaroo moving fastest?

Subject: at T0 (when the kangaroo takes off) (*correct*)

Independence of X and Y (KC32) – While motion in x and y is independent, that is we can think of it as two 1-D parts, they are each a function of time. An example from the stone problem's post-problem discussion:

System: If the stone is thrown with a larger horizontal initial velocity, what happens to the time spent in the air?

Subject: it stays in the air for the same amount of time (*correct*)

6.1.2.2 Pretest and Post-test There were two isomorphic tests, one randomly selected for the pretest and the other then used for the post-test. There were 48 multiple-choice questions on each test. The tests are adapted from two earlier studies, both on post-problem discussions with tutors [66, 65]. Questions were converted to multiple choice and some questions were modified to include graphics. To measure performance on both illustration questions and graph questions, isomorphic test questions were written, one set of questions relating to illustrations and the other set relating to graphs.

The pre-existing test questions were already tagged with KCs from the prior study that the questions came from. Some of the new test questions were isomorphs of existing test questions and so were given the same tags as the original questions. Other new questions were testing how well the subject could interpret graphs and illustrations and thus had no applicable knowledge component. Both tests can be found in Appendix [G](#).

6.1.2.3 Tutoring While the pilot study had two tutoring problems, one involving vertical motion and one involving projectile motion, this study had four problems in total. The first two were one-dimensional motion problems. The first problem was a horizontal motion problem involving a supertanker slowing to a stop. The second was a vertical motion problem involving a kangaroo jumping up. This kangaroo problem was a modified version of the kangaroo problem from the first study. The reflection dialogues, illustrations, and graphs were updated based on feedback from the first study.

The last two problems were projectile motion problems. The first involved an arrow being shot at a target and the second involved a stone being thrown off of a building. In both cases, the projectile was launched horizontally. The arrow problem was a modified version of the arrow problem from the first study. The reflection dialogues, illustrations, and graphs were updated based on feedback from the first study.

The arrow problem had three post-problem reflection dialogues; the other three had two. During tutoring, the representational form could change at the start of each problem and at the start of each reflection dialogue, giving a total of 13 points during tutoring during which a decision needed to be made on which representation to show. The problems and reflection questions can be seen in Appendix [H](#).

The problems, graphics, and reflection questions were approved by the high school physics teachers involved in the Rimac project. The physics problems came from the standard set of problems provided by the Andes physics tutoring system [109]. The reflection dialogues for the supertanker and stone problems were adapted from the Rimac project's reflection dialogues [65], modified to incorporate graphics into the discussion. The graphics were created specifically for this study and were approved by the physics teachers as being relevant for the problems and concepts being discussed, as being similar to illustrations and graphs

used in standard physics education material, and that both the illustrations and graphs portray similar information.

6.1.3 Procedure

The flow of the study is described in Chapter 3. To determine whether adapting the scheduling improves learning, the experiment compared three conditions (see below). Before tutoring, subjects were assigned to a condition. For the alternating and constant conditions, subjects in these conditions were then randomly assigned to one of the sub-conditions.

adaptive – the scheduling of graphs and illustrations adapts to the student and tutoring contexts using the $\text{PreScore} * \text{PctThruSession}$ policy learned in Chapter 5

alternating – the scheduling of graphs and illustrations alternate: at each point during tutoring, when a decision of which representation to show needs to be made, the representation switches; there are two sub-conditions based on which representation is shown first:

alternating-graph – the first representations shown are graphs

alternating-illus – the first representations shown are illustrations

constant – the representation shown remains constant throughout tutoring; there are two sub-conditions based on which representation is shown:

constant-graph – the representations shown are graphs

constant-illus – the representations shown are illustrations

6.1.4 Statistical Tests

In analysis of the results below, I make use of various statistical tests. In Section 4.1.4 I introduced independent-samples t-test and Pearson correlation, in this section I introduce the new tests.

The Analysis of Variance (ANOVA) is used to analyze the difference between group means. One or more categorical independent variables are used to determine the groups. If there is more than one independent variable, ANOVA can form groups using each independently, then also form groups considering values from both variables. By forming groups

Table 6.1: Distribution of students across conditions and sub-conditions.

Condition	Sub-condition	N
constant	graph	30
	illus	15
alternating	graph	15
	illus	14
adaptive		15
		27

using multiple variables, one can identify interactions between independent variables. If an interaction is found, a post-hoc test can be used to identify which groups are different from other groups. In this work, I use Tukey’s Honest Significant Difference (Tukey HSD) test.

6.2 RESULTS

Eighty-seven subjects without college physics were recruited and randomly assigned to one of the conditions. All 87 subjects showed learning gains. For one subject, in the alternating condition (alternating-graph sub-condition), I lost the logs for one problem. Therefore, in the analysis below, I will be using data from the remaining 86 subjects. Table 6.1 shows the distribution of students.

To verify that the conditions are balanced for background physics knowledge, I compare the pretest scores across the conditions with an ANOVA. With a p-value of 0.493, I confirm that the conditions are balanced for background physics knowledge. I also check that conditions are balanced for gender using the Pearson Chi-Squared test and age using an ANOVA and find that there are no significant differences across conditions ($p = 0.408$, $p = 0.409$ respectively). To check that the two tests were isomorphic, I ran an independent-samples t-test on the pretest scores and found no significant difference ($p = 0.610$).

Overall, subjects showed significant learning gains ($p = 0.000$) from pretest to post-test,

scoring 29.0 out of 48 on the pretest and 37.7 out of 48 on the post-test. Within each condition, subjects showed significant learning gains. For the constant condition, subjects scored 29.7 on the pretest and 37.0 on the post-test ($p = 0.000$). For the alternating condition, subjects scored 30.3 on the pretest and 39.4 on the post-test ($p = 0.000$). For the adaptive condition, subjects scored 28.7 on the pretest and 36.6 on the post-test ($p = 0.000$).

In the following sections, “Problem-Solving Graphics” will be abbreviated “PS” and “Post-Problem Discussion” will be abbreviated “PPD”.

6.2.1 Single Representation versus Multiple Representations

I first test the hypothesis that there is an advantage to showing multiple visual representations over showing a single visual representation in a natural language-based conceptual physics tutoring system. This is done by comparing two groups. The single representation group consists of the constant condition. The multiple representation group consists of the alternating and adaptive conditions. In Section 6.2.1.1, I examine differences in learning gains and satisfaction between the single representation group and the multiple representation group. I then examine interaction effects with measures of incoming knowledge and gender in Section 6.2.2.3. Section 6.2.1.3 investigates whether one group of students learned faster than another by examining learning curves.

6.2.1.1 Learning Gains and Satisfaction Table 6.2 compares learning gains and satisfaction across the two groups using independent samples t-test. Here I use raw learning gains, although similar results were found for normalized learning gains and post-test score controlled for pretest. For learning gains, we see that there is one significant difference, for KC32. Here, subjects seeing only one representation perform significantly better on KC32 than those seeing multiple representations. However, it is unclear whether this significant difference is from the number of representations seen or from the kind of graphic seen. We also see significant differences for satisfaction. For the three satisfaction measures involving graphics, we see that subjects seeing multiple representations preferred the graphics more than those who saw only one representation. While it may seem that this suggests subjects

prefer seeing multiple representations over a single representation, it is not clear whether multiple representations are preferred over only illustrations or only graphs.

To investigate, I use an ANOVA to compare constant-graph, constant-illus, and multiple representations; the results can be found in Table 6.3. We see that there are significant differences between the three representations for each significant measure of learning or satisfaction. Post-hoc comparisons using Tukey HSD can be found in Table 6.4. We see that there are significant differences between constant-graph and multiple representations for the satisfaction measures, with subjects preferring the graphics more when they see multiple representations. Comparing constant-illus and multiple representations, we see that subjects perform significantly better on KC32 when seeing only illustrations than when seeing multiple representations. Finally, comparing constant-graph to constant-illus, we see that subjects prefer the illustrations over graphs for both problem solving and in general. To conclude, while multiple representations may be better than graphs, it appears that illustrations are best overall, performing either better than or no different than graphs and multiple representations.

6.2.1.2 Interaction Effects The previous section suggests that learners should only see illustrations; showing multiple representations is either not significantly different or in some cases significantly worse than just showing illustrations. This result is consistent with some literature suggesting that illustrations are more beneficial for novices [81, 83, 57]. Since none of the subjects in this study have had college physics, they can be considered novices. However, since the best representation or scheduling can depend on a person's gender [84] and knowledge [70, 79], I suspect that there might be interaction effects between the number of representations seen and certain groupings of students.

Since the best representation may depend on knowledge or gender, I now investigate interaction effects between the number of representations seen and gender and knowledge. Incoming knowledge can be determined from the subject's score on the pretest, from the background survey question asking whether they had had high school physics, and from the background survey question asking whether they have had Calculus 1. For pretest score, a median split (median: 29.5/48) was performed to divide students into high and

Table 6.2: Comparing single representation to multiple representations across learning gains and satisfaction using independent samples t-test. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5. For all tests, the degrees of freedom are 84.

Measure Type	Measure	N	Single Representation (N = 30)	Multiple Representations (N = 56)	p-value	power
Learning Gain	Overall	48	9.07	8.54	0.6759	0.0699
	KC15	12	2.80	2.93	0.8719	0.0529
	KC23	12	0.40	0.48	0.5650	0.0881
	KC24	4	0.43	0.41	0.8939	0.0520
	KC25	2	0.50	0.55	0.6397	0.0751
	KC28	13	2.03	1.98	0.9230	0.0511
	KC32	3	0.73	0.46	0.0281	0.5982
Satisfaction	PS Graphics	4	4.17	4.50	0.0240	0.6224
	PPD	8	3.87	3.83	0.8272	0.0554
	PPD Graphics	6	4.07	4.47	0.0063	0.7917
	Overall System	5	3.67	3.76	0.6578	0.0724
	Graphics Satisfaction	10	4.12	4.48	0.0054	0.8060
	Overall Satisfaction	23	3.94	4.14	0.1285	0.3294

Table 6.3: Comparing constant-graph, constant-illus and multiple representations across the significant measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure	N	Constant-graph (n = 15)	Constant-illus (n = 15)	Multiple Repr. (n = 56)	$F_{2,84}$	p-value	power
KC32	3	0.6000	0.8667	0.4643	3.471	0.0356	0.424
PS Graphics	4	3.8167	4.5167	4.4955	7.931	0.0007	0.438
PPD Graphics	6	3.8556	4.2889	4.4673	5.930	0.0039	0.433
Graphics Satisfaction	10	3.8361	4.4028	4.4814	8.600	0.0004	0.443

Table 6.4: Post-hoc comparisons between constant-graph, constant-illus and multiple representations across the significant measures of learning and satisfaction. For each measure, the top value in the cell represents the difference of the average of the two groups. The middle value is the p-value and the bottom value is the power. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure	Constant-graph vs. Mult. Repr.	Constant-illus vs. Mult. Repr.	Constant-graph vs. Constant-illus
Degrees of Freedom	69	69	28
KC32	0.1357 0.653 0.131	0.4024 0.028 0.725	-0.2667 0.356 0.366
PS Graphics	-0.6789 0.001 0.957	0.0211 0.992 0.053	-0.7000 0.006 0.639
PPD Graphics	-0.6117 0.003 0.933	-0.1783 0.578 0.170	-0.4333 0.135 0.412
Graphics Satisfaction	-0.6453 0.000 0.983	-0.7862 0.870 0.085	-0.5667 0.014 0.637

low pretesters. The subject's gender was also recorded from the background survey. To determine whether there is an interaction effect, I perform a two-factor ANOVA, with single versus multiple representations as one factor and incoming knowledge or gender as the other.

Table 6.5 shows the interactions for pretest; none are significant. Table 6.6 shows the interactions for high school physics, where we do find a significant interaction with Post-Problem Discussion. We see that subjects who have had high school physics prefer the post-problem discussions while seeing only a single representation over seeing multiple representations. Those who have not had high school physics prefer post-problem discussions while seeing multiple representations over single representations.

Table 6.7 shows the interaction results for those who have had (or were currently taking) calculus 1. We see one significant interaction effect with KC25. Here, subjects who have not had calculus yet perform much better when seeing multiple representations than when seeing a single representation. Those with calculus 1 experience do better when seeing a

Table 6.5: Examining interaction effects between a median split of pretest score and number of representations seen. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure Type	Measure	N	$F_{2,84}$	p-val	power	Single		Multiple	
						High (n=12)	Low (n=18)	High (n=30)	Low (n=26)
Learning Gain	Overall	48	0.224	0.7995	0.225	6.9167	10.5000	7.1667	10.1154
	KC15	12	0.942	0.3940	0.322	3.1667	2.5556	2.5667	3.3462
	KC23	12	0.711	0.4940	0.312	0.5000	0.3333	0.4667	0.5000
	KC24	4	0.023	0.9774	0.058	0.1667	0.6111	0.2000	0.6538
	KC25	2	1.189	0.3095	0.327	0.2500	0.6667	0.4667	0.6538
	KC28	13	0.435	0.6487	0.283	1.3333	2.5000	1.6667	2.3462
	KC32	3	0.697	0.5008	0.311	0.7500	0.7222	0.4000	0.5385
Satisfaction	PS Graphics	4	1.049	0.3550	0.325	4.0417	4.2500	4.5250	4.4615
	PPD	8	3.103	0.0501	0.383	3.6354	4.0208	3.9417	3.7067
	PPD Graphics	6	2.816	0.0655	0.382	3.8333	4.2315	4.5278	4.3974
	Overall System	5	0.511	0.6020	0.294	3.6000	3.7111	3.8050	3.7000
	Graphics Satis.	10	2.155	0.1223	0.332	3.9375	4.2407	4.5264	4.4295
	Overall Satis.	23	2.219	0.1150	0.332	3.7776	4.0534	4.1999	4.0664

Table 6.6: Examining interaction effects between high school physics experience and number of representations seen. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure Type	Measure	N	$F_{2,84}$	p-val	power	Single		Multiple	
						Yes (n=17)	No (n=13)	Yes (n=41)	No (n=15)
Learning Gain	Overall	48	0.937	0.3960	0.322	8.4706	9.8462	8.7805	7.8667
	KC15	12	0.177	0.8380	0.202	2.6471	3.0000	2.9268	2.9333
	KC23	12	0.410	0.6651	0.278	0.2353	0.6154	0.4146	0.6667
	KC24	4	0.561	0.5730	0.300	0.4706	0.3846	0.4878	0.2000
	KC25	2	1.037	0.3590	0.324	0.5294	0.4615	0.5122	0.6667
	KC28	13	1.188	0.3100	0.327	1.7647	2.3846	2.1220	1.6000
	KC32	3	0.460	0.6327	0.287	0.7059	0.7692	0.4146	0.6000
Satisfaction	PS Graphics	4	1.030	0.3616	0.324	4.2353	4.0769	4.4634	4.5833
	PPD	8	3.240	0.0441	0.433	4.0588	3.6154	3.7744	3.9917
	PPD Graphics	6	2.215	0.1155	0.333	4.1373	3.9872	4.3821	4.7000
	Overall System	5	1.689	0.1909	0.331	3.9882	3.2462	3.8085	3.6133
	Graphics Satis.	10	1.840	0.1651	0.332	4.1863	4.0321	4.4228	4.6417
	Overall Satis.	23	2.773	0.0682	0.383	4.1049	3.7314	4.1071	4.2221

Table 6.7: Examining interaction effects between Calculus experience and number of representations seen. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure Type	Measure	N	$F_{2,84}$	p-val	power	Single		Multiple	
						Yes (n=11)	No (n=19)	Yes (n=26)	No (n=30)
Learning Gain	Overall	48	0.624	0.5730	0.306	9.2632	8.7273	9.4667	7.4615
	KC15	12	0.298	0.7430	0.253	2.7368	2.9091	3.1000	2.7308
	KC23	12	0.052	0.9490	0.099	0.3684	0.4545	0.4333	0.5385
	KC24	4	0.151	0.8600	0.187	0.3684	0.5455	0.3000	0.5385
	KC25	2	4.216	0.0180	0.494	0.5789	0.3636	0.4000	0.7308
	KC28	13	0.099	0.9060	0.148	2.1579	1.8182	2.2000	1.7308
	KC32	3	0.014	0.9859	0.041	0.7368	0.7273	0.4667	0.4615
Satisfaction	PS Graphics	4	0.753	0.4740	0.315	4.1053	4.2727	4.5167	4.4712
	PPD	8	1.020	0.3650	0.324	3.7895	4.0000	3.8708	3.7885
	PPD Graphics	6	1.584	0.2113	0.331	3.9386	4.3030	4.4667	4.4679
	Overall System	5	2.732	0.0709	0.411	3.5053	3.9455	3.9000	3.5904
	Graphics Satis.	10	1.322	0.2722	0.328	4.0219	4.2879	4.4917	4.4696
	Overall Satis.	23	2.149	0.1230	0.333	3.8346	4.1303	4.1885	4.0795

single representation.

The interaction results for Gender can be found in Table 6.8. We again see a significant interaction effect with KC25. Females perform better when seeing a single representation over seeing multiple representations. Males appear to perform better seeing multiple representations but since there are so few males in the single representation condition, this may not be the case.

When examining interaction effects, we found instances where some students benefit from multiple representations over a single representation. Subjects who have not had calculus have greater learning gains on KC25 when seeing multiple representations over seeing only one representation. The same may be true for males. Subjects who have not had high school physics prefer the post-problem discussions when seeing multiple representations over seeing a single representation. Thus, it appears that subjects who have not had calculus and subjects who have not had physics both benefit from multiple representations. This is in contrast to the findings in Section 5.3.1, where the model predicted showing low pretesters a

Table 6.8: Examining interaction effects between gender and number of representations seen. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure Type	Measure	N	$F_{2,84}$	p-val	power	Single		Multiple	
						Female (n=23)	Male (n=7)	Female (n=35)	Male (n=21)
Learning Gain	Overall	48	0.176	0.8393	0.239	9.7391	6.8571	9.4000	7.0952
	KC15	12	0.007	0.9930	0.062	3.0435	2.0000	3.3143	2.2857
	KC23	12	0.681	0.5090	0.227	0.4348	0.2857	0.4571	0.5238
	KC24	4	1.102	0.3371	0.328	0.5652	0.0000	0.4857	0.2857
	KC25	2	4.318	0.0164	0.433	0.6522	0.0000	0.5714	0.5238
	KC28	13	0.235	0.7910	0.260	2.2609	1.2857	2.2286	1.5714
	KC32	3	0.916	0.4042	0.325	0.7826	0.5714	0.4571	0.4762
Satisfaction	PS Graphics	4	0.221	0.8018	0.256	4.1413	4.2500	4.4857	4.5119
	PPD	8	0.399	0.6720	0.294	3.8750	3.8393	3.7893	3.9048
	PPD Graphics	6	1.473	0.2351	0.331	4.1522	3.8095	4.4524	4.4921
	Overall System	5	1.105	0.3360	0.328	3.7478	3.4000	3.7200	3.8167
	Graphics Satis.	10	0.503	0.6066	0.327	4.1467	4.0298	4.4690	4.5020
	Overall Satis.	23	0.824	0.4420	0.323	3.9791	3.8247	4.1118	4.1813

single representation (illustrations) while high pretesters should see multiple representations. This difference in results may be due to different students, different sample sizes, or changes in images or tutoring environment. It is also possible that pretest score measures something similar, but different, than high school physics experience, such as physics knowledge versus experience with physics. A similar possibility exists with calculus 1 experience, which could measure whether students have been taught to think a certain way versus physics knowledge.

6.2.1.3 Learning Curves In the previous section, I examined learning gains between groups of subjects. Learning gains are computed from a test given before the tutoring and another test given after. Another option for measuring learning is to use learning curves. A learning curve is a plot of the error on a knowledge component during tutoring over the number of opportunities to practice that knowledge component [80]. If a subject is learning, then as the number of opportunities increase, the error rate should decrease and thus the learning curve should decrease. The error rates for each KC can be aggregated for each student, giving the performance of any skill over the opportunities to practice it [80]. A

curve can be fit to the data, giving an equation representing the relationship between error rate and number of opportunities to practice. Learning curves have been observed to follow a power law relation shown in equation 6.1 [80, 85]:

$$Y = aX^b \tag{6.1}$$

where:

Y is the error rate

X is the number of opportunities to practice the knowledge component

The parameters a and b are learned as the curve is fit to the data. The parameter a is the error rate of the first opportunity (representing the intrinsic difficulty of the KC), and b is the learning rate. Negative learning rates indicate that the subject is making fewer errors over the opportunities. A larger magnitude for b indicates a greater rate of change in the error rate over the opportunities. Finally, the fit of the curve (adjusted r^2) can be a measure of how much learning is taking place with respect to the task being measured [80].

Following the suggestions of the literature [80], I aggregate the KC error rates for individual subjects, fit a power curve to the aggregated error rate for each subject, then compare the learning rate (b) and the adjusted r^2 value across subjects. One open question is how many opportunities to include [80]. As the number of opportunities increases, the number of KCs still in use tends to decrease. Some use every opportunity [22, 67] while other suggest establishing a threshold such as when the number of data points in an opportunity drops below 50% of the number in the first opportunity [80]. Since there is no standard threshold to use, I experimented with both. I found that no threshold generally produced more interesting results for comparing single versus multiple representation learning curves. For alternating versus adaptive learning curves, I found that the 50% threshold produced more interesting curves. In future work, it would be good to develop a more formal method for determining the threshold a priori.

The average learning rate for subjects seeing a single representation is -1.070 and the adjusted r^2 value is 0.240. For subjects seeing multiple representations the learning rate is -1.167 and the adjusted r^2 is 0.293. Using an independent samples t-test, we find that there

Table 6.9: Interaction effect results for pretest and single vs. multiple representations on learning curves.

measure	p-val	single		multiple	
		High	Low	High	Low
learning rate	0.1210	-1.1690	-1.0225	-1.0895	-1.2518
Adj. r^2	0.1750	0.2185	0.1530	0.2131	0.2527

Table 6.10: Interaction effect results for high school physics experience and single vs. multiple representations on learning curves.

measure	p-val	single		multiple	
		No	Yes	No	Yes
learning rate	0.6830	-1.0403	-1.1122	-1.1757	-1.1626
Adj. r^2	0.8590	0.1538	0.1986	0.2097	0.2402

is no significant difference between the two groups of subjects on learning rate ($p = 0.310$) or adjusted r^2 ($p = 0.275$).

As with the learning gains and satisfaction analysis above, I now examine whether there are interaction effects with incoming knowledge and gender. For pretest (Table 6.9), high school physics experience (Table 6.10), and gender (Table 6.12), we see no significant differences. For Calculus, we find significant interaction effects on both learning rate and adjusted r^2 . For both, subjects who have had calculus learn faster and have more of their learning explained by the number of opportunities they have had when seeing single representations over seeing multiple representations. The opposite occurs for subjects who did not have calculus; they do better when seeing multiple representations. We again find that those without calculus perform better with multiple representations.

6.2.2 Alternating Representations vs. Adaptive Scheduling of Representations

I now test the hypothesis that adaptive scheduling of representations is more beneficial than alternating representations in a natural language-based conceptual physics tutoring system.

Table 6.11: Interaction effect results for calculus experience and single vs. multiple representations on learning curves.

measure	p-val	single		multiple	
		Yes	No	Yes	No
learning rate	0.0024	-1.2223	-0.8371	-1.0683	-1.2837
Adj. r^2	0.0144	0.2309	0.0898	0.2100	0.2580

Table 6.12: Interaction effect results for gender and single vs. multiple representations on learning curves.

measure	p-val	single		multiple	
		Female	Male	Female	Male
learning rate	0.2760	-1.0248	-1.2658	-1.1670	-1.1648
Adj. r^2	0.3330	0.1585	0.2471	0.2301	0.2349

In Section 6.2.2.1 the frequency of switching for alternating and adaptive policies are compared, found to be different, and some investigations into how frequency of switching affects learning are performed. In Section 6.2.2.2, I examine differences in learning gains and satisfaction between subjects who saw representations according to a fixed, alternating schedule and subjects who saw representations according to an adaptive schedule. I then examine interaction effects with measures of incoming knowledge and gender in Section 6.2.2.3. Section 6.2.2.4 investigates whether one group of subjects learned faster than another by examining learning curves.

6.2.2.1 Frequency of Switching Table 6.13 shows the frequency of switching for each of the subjects. From this table, we see that the adaptive condition switched representations between one and four times; the alternating condition switched 12 times.

As seen below, the alternating condition often performs no differently than the adaptive condition. One possible reason for this is that the alternating condition does occasionally show the same graphic as the adaptive model would predict. Does the alternating condition do no different than adaptive because it shows the same graphic as the adaptive model

Table 6.13: Number of decision points during tutoring where the representation changed.

# Switches	# Subjects	# Alternating	# Adaptive
1	15	0	15
3	11	0	11
4	1	0	1
12	29	29	0

Table 6.14: Distribution of how often the alternating scheduling picked the same representation that the adaptive policy would have picked.

# Matches	# Subjects
4	9
6	3
7	5
8	12

predicts? To answer this question, I first ran the data from the alternating subjects through the adaptive model to get the model’s predictions. I counted the number of times the graphic type the model predicted agreed with what the subject saw. Table 6.14 is a summary of the results. Note that there are a total of 13 places in tutoring when the tutor decided which type to show next.

There is a plurality of subjects who saw the graphic suggested by the adaptive model 61% (8/13) of the time. This suggests that it is possible that alternating did so well because it fairly often did what adaptive would do. If that’s the case, then we would expect a positive correlation between the number of matches and learning or satisfaction. I ran Pearson correlations to test this. From Table 6.15, we see no correlation between number of matches and learning or satisfaction. These results fail to support the hypothesis that alternating did so well because it mimicked the adaptive condition.

Based on the analysis so far, it is not clear why alternating does as well as adaptive. It could be that frequent switching produces learning gains as good as adapting to the user. It could also be that the adaptive policy needs to be improved. If frequent switching does

Table 6.15: Correlations between number of times the alternating condition agreed with the adaptive model’s prediction and learning or satisfaction.

Measure Type	Measure	N	r	p
Learning Gain	Overall	48	0.1143	0.5550
	KC15	12	-0.0306	0.8747
	KC23	12	-0.0880	0.6497
	KC24	4	0.2110	0.2718
	KC25	2	-0.2308	0.2285
	KC28	13	0.2646	0.1653
	KC32	3	0.0390	0.8407
Satisfaction	PS Graphics	4	-0.3096	0.1022
	PPD	8	-0.1902	0.3230
	PPD Graphics	6	-0.2489	0.1930
	Overall System	5	-0.2421	0.2058
	Graphics Satisfaction	10	-0.3134	0.0978
	Overall Satisfaction	23	-0.2982	0.1162

improve learning, then we would expect more frequent switches to positively correlate with learning. To evaluate this, I examined the correlations between number of times the representations switched and learning gains and satisfaction on the alternating and adaptive conditions. From Table 6.16, we see no such correlations. However, when examining the correlations just within the adaptive condition (see Table 6.17), we find that there is a significant negative correlation for KC24. This suggests that for KC24 (gravitational acceleration), more frequent switching adversely affects learning. This does not support the hypothesis that frequent switching positively correlates with learning. Instead, the model seems to have found schedules of graphic types that produces learning with fewer switches. This suggests that the adaptive policy is able to produce learning gains without many switches.

From the analysis in this section, we fail to find evidence suggesting that the frequency of switching can account for differences in learning between the alternating and adaptive conditions. While the literature does suggest that frequent switching is better than infrequent switching [91], the subjects are different. In future work it would be good to control for the number of switches subjects see.

Table 6.16: Correlations between number of times representations switched and learning or satisfaction, on alternating and adaptive conditions.

Measure Type	Measure	N	r	p
Learning Gain	Overall	48	0.1390	0.3069
	KC15	12	-0.0574	0.6742
	KC23	12	-0.1413	0.2988
	KC24	4	0.1036	0.4475
	KC25	2	-0.1676	0.2170
	KC28	13	0.0103	0.9400
	KC32	3	0.2128	0.1153
Satisfaction	PS Graphics	4	0.0156	0.9090
	PPD	8	-0.1030	0.4499
	PPD Graphics	6	-0.0107	0.9376
	Overall System	5	0.0010	0.9943
	Graphics Satisfaction	10	0.0011	0.9936
	Overall Satisfaction	23	-0.0330	0.8094

Table 6.17: Correlations between number of times representations switched and learning or satisfaction, on the adaptive condition.

Measure Type	Measure	N	r	p
Learning Gain	Overall	48	-0.1138	0.5720
	KC15	12	-0.3164	0.1079
	KC23	12	-0.2418	0.2244
	KC24	4	-0.3963	0.0407
	KC25	2	0.0189	0.9254
	KC28	13	-0.0010	0.9959
	KC32	3	-0.1350	0.5018
Satisfaction	PS Graphics	4	0.2010	0.3147
	PPD	8	0.1486	0.4594
	PPD Graphics	6	0.2725	0.1690
	Overall System	5	0.1936	0.3334
	Graphics Satisfaction	10	0.2590	0.1920
	Overall Satisfaction	23	0.2360	0.2361

Table 6.18: Comparing adaptive scheduling to fixed scheduling of multiple representations across measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5. For all tests, the degrees of freedom are 54.

Measure Type	Measure	N	Adaptive (n=27)	Alternating (n=29)	p-value	power
Learning Gain	Overall	48	7.85	8.54	0.5291	0.096
	KC15	12	3.04	2.93	0.8847	0.052
	KC23	12	0.56	0.48	0.6119	0.079
	KC24	4	0.37	0.41	0.8301	0.055
	KC25	2	0.63	0.55	0.5168	0.099
	KC28	13	1.96	1.98	0.9741	0.050
	KC32	3	0.33	0.46	0.3440	0.156
Satisfaction	PS Graphics	4	4.49	4.50	0.9657	0.050
	PPD	8	3.92	3.83	0.6062	0.080
	PPD Graphics	6	4.48	4.47	0.9518	0.050
	Overall System	5	3.77	3.76	0.9549	0.050
	Graphics Satisfaction	10	4.48	4.48	0.9882	0.050
	Overall Satisfaction	23	4.16	4.14	0.8422	0.054

6.2.2.2 Learning Gains and Satisfaction Table 6.18 compares learning gains and satisfaction across the two conditions using an independent samples t-test. Here I use raw learning gains, although similar results were found for normalized learning gains and post-test score controlled for pretest. There are no significant differences across conditions. This might suggest that the frequency of switching might not be important when using multiple representations. This is contrary to prior research on the frequency of switching between multiple representations for a fractions tutor, which found that switching frequently is more beneficial than infrequent switching [91]. There are many differences between the fractions tutor study and the current one, including domain, subject population, and number of representations available. Future work is needed to investigate how frequency of switching graphical representations affects conceptual physics learning.

6.2.2.3 Interaction Effects I now investigate interaction effects to identify sets of students who benefit from an adaptive schedule over an alternating schedule. As with the single versus multiple representation analysis above, I consider three measures of knowledge and

Table 6.19: Interaction effects between a median split of pretest score and scheduling of multiple representations for measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure Type	Measure	N	$F_{1,55}$	p-val	power	Alternating		Adaptive	
						High (n=15)	Low (n=14)	High (n=9)	Low (n=18)
Learning Gain	Overall	48	1.464	0.2314	0.327	7.0000	11.5000	6.8889	8.3333
	KC15	12	2.594	0.1130	0.333	1.4667	4.2857	3.0000	3.0556
	KC23	12	0.155	0.6950	0.184	0.2667	0.5714	0.4444	0.6111
	KC24	4	0.168	0.6832	0.191	0.2667	0.6429	0.0000	0.5556
	KC25	2	4.184	0.0456	0.494	0.5333	0.4286	0.3333	0.7778
	KC28	13	0.031	0.8610	0.073	1.6667	2.3571	1.6667	2.1111
	KC32	3	0.629	0.4310	0.298	0.4667	0.7143	0.3333	0.3333
Satisfaction	PS Graphics	4	0.006	0.9370	0.027	4.5167	4.4821	4.5278	4.4722
	PPD	8	2.997	0.0890	0.409	3.9667	3.5179	3.7639	4.0000
	PPD Graphics	6	0.527	0.4710	0.287	4.5667	4.3452	4.4630	4.4815
	Overall System	5	6.443	0.0140	0.494	4.0400	3.4286	3.3556	3.9750
	Graphics Satis.	10	0.170	0.6820	0.192	4.5417	4.4137	4.4954	4.4769
	Overall Satis.	23	3.099	0.0839	0.411	4.2725	3.9435	4.0275	4.2322

gender.

The interaction results for pretest score can be found in Table 6.19. Here we see significant interaction effects for KC 25 and Overall System Satisfaction. On KC 25, low pretesters perform better when receiving an adaptive scheduling of representations than with fixed scheduling. High pretesters perform better on KC 25 with the fixed alternating scheduling than with the adaptive scheduling. We see similar results with Overall System satisfaction: low pretesters prefer the system when receiving an adaptive scheduling while high pretesters prefer the system when receiving the alternating scheduling.

The interaction results for high school physics are shown in Table 6.20. Here we have two significant interaction effects, one with KC 15 and the other with KC 23. For both KCs, subjects who had high school physics perform better when receiving an adaptive scheduling of representations than when receiving an alternating scheduling. Those who did not have high school physics perform better with an alternating scheduling.

Both high school physics and pretest score are intended to be measures of incoming

Table 6.20: Examining interaction effects between high school physics experience and scheduling of multiple representations for measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure Type	Measure	N	$F_{1,55}$	p-val	power	Alternating		Adaptive	
						Yes (n=23)	No (n=6)	Yes (n=18)	No (n=9)
Learning Gain	Overall	48	4.013	0.0501	0.383	8.6522	11.167	8.9444	5.6667
	KC15	12	6.854	0.0114	0.445	2.2609	5.0000	3.7778	1.5556
	KC23	12	13.263	0.0006	0.408	0.2174	1.1667	0.6667	0.3333
	KC24	4	1.054	0.3090	0.332	0.5652	0.0000	0.3889	0.3333
	KC25	2	3.360	0.0722	0.359	0.3913	0.8333	0.6667	0.5556
	KC28	13	0.131	0.7190	0.172	2.1739	1.3333	2.0556	1.7778
	KC32	3	0.454	0.5031	0.283	0.5652	0.6667	0.2222	0.5556
Satisfaction	PS Graphics	4	0.729	0.3970	0.315	4.4457	4.7083	4.4861	4.5000
	PPD	8	0.777	0.3820	0.318	3.7554	3.7292	3.7986	4.1667
	PPD Graphics	6	0.048	0.8271	0.097	4.3841	4.7500	4.3796	4.6667
	Overall Sys.	5	0.005	0.9460	0.025	3.7913	3.5667	3.8306	3.6444
	Graphics Satis.	10	0.320	0.5740	0.245	4.4149	4.7292	4.4329	4.5833
	Overall Satis.	23	0.006	0.9400	0.028	4.0941	4.1885	4.1237	4.2444

knowledge. We would expect to see similar results for both, but instead we appear to be finding contradictory results. For incoming knowledge according to the pretest, high knowledge does better with an alternating schedule and low knowledge does better with an adaptive schedule. For incoming knowledge according to high school physics experience, high knowledge does better with an adaptive schedule and low knowledge does better with an alternating schedule. These seemingly different results may be due to differences in the KCs. KC 15 and KC 23 both involve acceleration while KC 25 involves the horizontal component of velocity in projectile motion. It might be that with these subjects, acceleration is better with one scheduling method and projectile motion or velocity is better with a different schedule. However, there may be other reasons for the observed difference. Future research is needed to investigate.

There were no significant interaction results for calculus (Table 6.21) or gender (Table 6.22).

Table 6.21: Examining interaction effects between Calculus experience and scheduling of multiple representations for measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure Type	Measure	N	$F_{1,55}$	p-val	power	Alternating		Adaptive	
						Yes (n=18)	No (n=11)	Yes (n=12)	No (n=15)
Learning Gain	Overall	48	0.534	0.4680	0.295	10.2222	7.4545	8.3333	7.4667
	KC15	12	0.043	0.8370	0.091	3.0556	2.4545	3.1667	2.9333
	KC23	12	0.564	0.4560	0.299	0.3333	0.5455	0.5833	0.5333
	KC24	4	3.484	0.0673	0.360	0.5000	0.3636	0.0000	0.6667
	KC25	2	0.266	0.6078	0.238	0.3889	0.6364	0.4167	0.8000
	KC28	13	0.139	0.7110	0.178	2.2778	1.5455	2.0833	1.8667
	KC32	3	0.483	0.4900	0.288	0.6111	0.5455	0.2500	0.4000
Satisfaction	PS Graphics	4	0.250	0.6190	0.232	4.5417	4.4318	4.4792	4.5000
	PPD	8	2.799	0.1000	0.337	3.9167	3.4773	3.8021	4.0167
	PPD Graphics	6	0.037	0.8480	0.083	4.4722	4.4394	4.4583	4.4889
	Overall System	5	1.737	0.1930	0.346	3.9889	3.3455	3.7667	3.7700
	Graphics Satis.	10	0.136	0.7140	0.176	4.5069	4.4356	4.4688	4.4944
	Overall Satis.	23	1.526	0.2220	0.343	4.2299	3.9235	4.1266	4.1939

Table 6.22: Examining interaction effects between gender and scheduling of multiple representations for measures of learning and satisfaction. Learning gains range from 0 to the number of questions. Satisfactions range from 1 to 5.

Measure Type	Measure	N	$F_{1,55}$	p-val	power	Alternating		Adaptive	
						Female (n=18)	Male (n=11)	Female (n=17)	Male (n=10)
Learning Gain	Overall	48	0.000	0.9932	0.003	10.0556	7.7273	8.7059	6.4000
	KC15	12	0.024	0.8780	0.064	3.1667	2.2727	3.4706	2.3000
	KC23	12	0.000	0.9890	0.003	0.3889	0.4545	0.5294	0.6000
	KC24	4	0.088	0.7680	0.139	0.5000	0.3636	0.4706	0.2000
	KC25	2	0.000	0.9950	0.004	0.5000	0.4545	0.6471	0.6000
	KC28	13	0.986	0.3250	0.329	2.5000	1.1818	1.9412	2.0000
	KC32	3	0.295	0.5890	0.247	0.6111	0.5455	0.2941	0.4000
Satisfaction	PS Graphics	4	0.434	0.5130	0.280	4.4583	4.5682	4.5147	4.4500
	PPD	8	0.001	0.9700	0.012	3.7083	3.8182	3.8750	4.0000
	PPD Graphics	6	0.095	0.7590	0.145	4.4259	4.5152	4.4804	4.4667
	Overall System	5	0.042	0.8380	0.090	3.6889	3.8364	3.7529	3.7950
	Graphics Satis.	10	0.271	0.6050	0.239	4.4421	4.5417	4.4975	4.4583
	Overall Satis.	23	0.087	0.7690	0.138	4.0704	4.1845	4.1558	4.1779

Table 6.23: Interaction effect results for pretest and alternating vs. adaptive scheduling of representations on learning curves.

measure	p-val	alternate		adaptive	
		High	Low	High	Low
learning rate	0.0342	0.0726	-0.5321	-0.4262	-0.3532
Adj. r^2	0.7300	-0.0335	0.0325	-0.0384	-0.0190

6.2.2.4 Learning Curves In this section, I create learning curves to measure learning rate of the subjects and to evaluate how well the error rate is described by the number of opportunities to practice (adjusted r^2). As described in Section 6.2.1.3, I aggregate the KC error rates for individual subjects, fit a power curve to the aggregated error rate for each subject, then compare the learning rate and the adjusted r^2 value across subjects. In this section, I only consider the first n opportunities, where opportunity $n+1$ has fewer than 50% of the data points that opportunity 1 had.

The average learning rate for scheduling representations according to an alternating schedule is -0.187 and the adjusted r^2 value is -0.005. For subjects seeing representations according to an adaptive schedule, the learning rate is -0.388 and the adjusted r^2 is -0.028. Using an independent samples t-test, we find that there is no significant difference between the two groups of subjects on learning rate ($p = 0.221$) or adjusted r^2 ($p = 0.726$). Since no significant differences are found, I examine interaction effects with measures of incoming knowledge and with gender.

High school physics experience (Table 6.24) and gender (Table 6.26) show no significant interaction effects. However, we find significant interactions with pretest (Table 6.23) and with calculus (Table 6.25). For pretest, we find that high pretesters learn more quickly when receiving an adaptive scheduling, whereas low pretesters learn more quickly when receiving an alternating scheduling.

Significant interactions also exist with calculus (Table 6.25). On learning rate, subjects with calculus learn much quicker with the adaptive model whereas subjects without calculus learn quicker with the alternating model. On adjusted r^2 , subjects who had calculus have

Table 6.24: Interaction effect results for high school physics experience and alternating vs. adaptive scheduling of representations on learning curves.

measure	p-val	alternate		adaptive	
		Yes	No	Yes	No
learning rate	0.0769	-0.0417	-0.7178	-0.3781	-0.4088
Adj. r^2	0.4010	-0.0097	0.0112	0.0074	-0.0997

Table 6.25: Interaction effect results for calculus experience and alternating versus adaptive scheduling of representations on learning curves.

measure	p-val	alternate		adaptive	
		Yes	No	Yes	No
learning rate	0.0220	-0.0614	-0.4119	-0.6177	-0.2048
Adj. r^2	0.0164	-0.0732	0.1172	0.0445	-0.0866

better fits with the adaptive model while subjects who did not have better fits with the alternating model.

Overall, it appears that subjects with some incoming experience do better with the adaptive model whereas subjects with low incoming experience do better with the alternating mode. It might be that subjects with little incoming experience learn more quickly when the representations are frequently switching, which is consistent with the literature [89, 91], or that the adaptive model did not perform well for the low pretesters. In Section 6.2.3 I explore improving the model.

Table 6.26: Interaction effect results for gender and alternating vs. adaptive scheduling of representations on learning curves.

measure	p-val	alternate		adaptive	
		Female	Male	Female	Male
learning rate	0.1300	-0.2467	-0.0783	-0.2588	-0.6084
Adj. r^2	0.2990	0.0573	-0.1177	-0.0158	-0.0496

6.2.3 Modeling Revisited

From the results above, we see instances when an adaptive scheduling is better than an alternating scheduling, but other times when there is no significant difference or even times when alternating does better. One potential reason for this is that the adaptive policy may not be very good. The best model identified in Chapter 5 was built from limited data. The sample size was small, only 2043 data points. Additionally, subjects in the first study only ever saw one representation during tutoring whereas in this study subjects could see two. Therefore, the model was not able to consider the effects of a subject having seen or not having seen a representation before when deciding which representation to use. Additionally, the number of problems increased and the two problems that were used in both studies were modified before being used in the new study. These limitations may explain why the adaptive condition did not consistently perform better than alternating.

To start investigating the claim that the adaptive model was limited, I use the method described in Section 5.2 to learn a new model from the data collected in this study. Additionally, this new model will also have access to new features that capture information related to subjects potentially seeing multiple representations. Once this new model is learned, I show that it performs better than two baselines on an intrinsic evaluation. In future work, I propose running an experiment similar to the one performed in Chapter 5 to learn the best adaptive policy given the data available.

In addition to the features described in Section 5.1, I add the new features shown below. Some have been median split; Table 6.27 shows a distribution of the values for the variables that were median split.

RepSeenPrior – Has the subject seen the current representation before the most recent decision point (no, yes)

RepLastSeenSec – Number of seconds since the subject last saw the representation (median split: high, low)

RepLastSeenTurn – Number of dialogue turns since the subject last saw the representation (median split: high, low)

RepTimeSpentSec – Total number of seconds the subject spent with the representation

Table 6.27: Low, median, and high values for the new representation features that were median split.

Feature	Low	Median	High
RepLastSeenSec	0.91	200.28	1844.35
RepLastSeenTurn	7.00	52.00	313.00
RepTimeSpentSec	0.00	1684.87	9143.52
RepTimeSpentTurns	0.00	264.00	1188.00
RepPretest	3.00	10.00	17.00
RepPerformance	0.22	0.69	1.00
RepPerformancePrev	0.00	1.00	1.00
RepSwitches	0.00	1.00	12.00

(median split: high, low)

RepTimeSpentTurns – Total number of turns the subject spent with the representation

(median split: high, low)

RepPretest – The percent correct on pretest questions involving the representation (median

split: high, low)

RepPerformance – The percent correct of dialogue turns (prior to the current dialogue)

involving the representation (median split: high, low)

RepPerformancePrev – The percent correct on dialogue turns from the most recent pre-

vious dialogue that involving the representation (median split: high, low)

RepSwitches – How frequent the representation had been switching; the ratio of number of

times the representation switched over the number of opportunities for the representation to switch (median split: high, low)

To learn the new model, I follow the steps described in Section 5.2. Since I am currently only interested in showing that a better model exists, I learn the simplest model: the **1 Factor** model. The 1 Factor model has only one factor in each term describing the tutoring context, plus a second factor indicating the representation to show. The initial equation is

of the form:

$$\begin{aligned}
 postscore = & prescore \\
 & + (RepIsGraph + RepIsIllus) * \sum_{factor_i \in F_1} factor_i
 \end{aligned}$$

where RepIsGraph and RepIsIllus are binary factors indicating whether the subject saw a graph or saw an illustration, and F_1 is the set of all factors except RepIsGraph and RepIsIllus.

The 1 Factor model is compared to two baselines. The first baseline is similar to the baseline from Chapter 5. This **Single-Graphic Baseline** always predicts showing the same representation. I use stepwise regression, starting with the equation $postscore = prescore + RepIsGraph + RepIsIllus$, to identify which representation to always show. Stepwise regression eliminates RepIsGraph, leaving RepIsIllus with a positive coefficient, indicating that illustrations should be the graphic to always show. Thus, the baseline used here is the same baseline from Chapter 5.

The second baseline comes from the PreScore*Gender model identified as the best from Chapter 5. Thus, I call it the **PreScore*Gender Baseline**. I take the terms in the PreScore*Gender model from Section 5.3.2, form a new equation, and perform linear regression with the data from the new study to learn new coefficients. By doing this, I am examining whether the features transfer rather than the entire model.

As in Chapter 5, I perform 10-fold cross-validation for each of the three models described above and compare the adjusted r^2 values. All subjects who completed the study were included in this analysis. Table 6.28 shows the results.

The 1 Factor model performs significantly better than the two baselines, suggesting that the additional features benefit modeling. Below are the rules extracted from the 1 Factor model, ranked according to the magnitude of the coefficient (coefficient shown in parentheses; larger indicates more important; the new features are bolded):

1. If HSPHysics = No, show Illus (2.6926)
2. If SessionPctCorrect = High, show Graph (2.5050)
3. If **RepPerformance = High, show Graph (2.3797)**

Table 6.28: Models learned are compared according to adjusted r^2 values and their 95% confidence intervals.

Model	Adj. r^2	95% CI
Single-Graphic Baseline	0.514	(0.505, 0.524)
PreScore*Gender Baseline	0.557	(0.548, 0.565)
1 Factor Model	0.601	(0.594, 0.607)

4. If ProblemPctCorrect = High, show Illus (2.1863)
5. **If RepPretest = High, show Graph (2.0186)**
6. If PctThruSession = Late, show Graph (1.6913)
7. **If RepSwitches = Frequent, show Illus (1.6825)**
8. If Gender = Male, show Illus (1.3350)
9. **If RepTimeSpentTurns = High, show Illus (1.1616)**
10. If RQPctCorrect = High, show Graph (0.9148)
11. If PctThruProblem = Late, show Graph (0.7793)
12. **If RepTimeSpentSec = High, show Illus (0.7377)**
13. If TimeThruSession = Late, show Illus (0.7234)
14. If PctThruProblem = Early, show Illus (0.6413)
15. If DialogStage = ProblemSolving, show Graph (0.6329)
16. **If RepSeenPrior = True, show Illus (0.5952)**
17. **If RepLastSeenSec = High, show Illus (0.5348)**
18. If PctThruSession = Early, show Illus (0.5181)
19. If KCusage = Apply, show Illus (0.3123)
20. If ResponseTimePrevWords = Low, show Graph (0.2580)

The new representation features (bolded above) appear to be important, appearing in seven of the twenty rules and four being in the top half of the rules. For example, the third most important feature is RepPerformance, which says that if learners do well when seeing graphs, they should continue to see graphs. The fifth most important (RepPretest) similarly says that if learners do well on graph pretest questions, they should see graphs

during tutoring. These findings are consistent with the literature which suggests that since abstract representations of the tutoring context, such as graphs, are believed to be better for students who can handle them [57, 70], learners who are doing well with graphs should see them.

There also appears to be a preference for showing graphs when subjects show evidence of understanding the concepts and illustrations when they show evidence of not understanding concepts, which is consistent with the literature [69, 57, 70]. When `SessionPctCorrect` or `RQPctCorrect` are High, subjects should see graphs. If the subject has not had high school physics, then they should see illustrations. One exception exists for this pattern. If the subject has been correct frequently within a problem, they should see illustrations. It is not clear why many correct answers within a problem suggests seeing illustrations, but many correct answers in the tutoring session or in a reflection dialogue suggest graphs. Adding more contextual factors may provide insights.

Additionally, we are finding rules that suggest subjects early in tutoring should see illustrations and subjects later in tutoring should see graphs. This is also consistent with the literature, which suggests that learners do best when transitioning from concrete representations (e.g. illustrations) to more abstract representations (e.g. graphs) [57, 83]. When `PctThruSession` is Late or when `PctThruProblem` is High, subjects should see graphs. When `PctThruSession` is Early or when `PctThruProblem` is Early, subjects should see illustrations.

To summarize, we found that the 1 Factor model may perform better than the baselines because of the representation features that have been added. Almost half of the 20 rules include these new features and almost half of the top 10 most important rules include these features. The model produces rules that are consistent with the literature. Subjects who can handle graphs or subjects who are doing well should see graphs. There should also be a general trend over time from showing illustrations to showing graphs.

We also observe that the `PreScore*Gender` baseline performs significantly better than the Single-Graphic Baseline, which suggests that the modeling approach might produce models that can transfer from one situation to another, although the weights might need to be relearned. Table 6.29 shows rules for this model. The rules learned with this new dataset were generally similar to those learned from the first study. Of the 20 rules, 8 switched their

conclusion (e.g. SessionPctCorrect=High for Female High Pretesters was originally Illus, but when relearning it becomes Graph). Since the sign of the coefficient can indicate which representation to show, a conclusion may change if the sign of the coefficient changed. The ranking of the rules can shift as coefficients change. On average, the rules did not shift much, shifting up or down only 1.3 positions.

Table 6.29: Rules for the PreScore*Gender baseline.

Female High Pretesters (n = 20)	Female Low Pretesters (n = 38)
<ol style="list-style-type: none"> 1. If ProblemPctCorrect=High, show Illus 2. If PctThruSession=Early, show Graph 3. If SessionPctCorrect=High, show Graph 4. If PctThruProblem=Early, show Graph 5. If WalkThruPctCorrect=Low, show Graph 6. If RQPctCorrect=Low, show Illus 	<ol style="list-style-type: none"> 1. If SessionPctCorrect=High, show Graph 2. If ProblemPctCorrect=High, show Illus 3. If RQPctCorrect=Low, show Graph 4. If PctThruSession=Early, show Illus 5. If PctThruProblem=Early, show Illus
Male High Pretesters (n = 22)	Male Low Pretesters (n = 6)
<ol style="list-style-type: none"> 1. If SessionPctCorrect=High, show Illus 2. If RQPctCorrect=Low, show Illus 3. If WalkThruPctCorrect=Low, show Illus 	<ol style="list-style-type: none"> 1. If RQPctCorrect=Low, show Graph 2. If WalkThruPctCorrect=Low, show Graph 3. If SessionPctCorrect=High, show Illus 4. If PctThruSession=Early, show Illus 5. If ProblemPctCorrect=High, show Graph 6. If PctThruProblem=Early, show Illus

6.3 DISCUSSION

The study presented in this chapter presented illustrations and graphs to subjects according to one of three schedules during conceptual physics tutoring. One schedule, the constant schedule, showed only illustrations or only graphs during tutoring. Another schedule, the alternating schedule, switched representations at the start of each dialogue. The third schedule, the adaptive schedule, consulted the adaptive policy learned in Chapter 5 to decide which representation to show at the start of each dialogue. A summary of findings from this chapter can be found in Table 6.30

One question I wished to address with this study was whether multiple representations are helpful in dialogue-based conceptual physics tutoring. In answering this question, I compared subjects in the constant schedule condition to the subjects in the other two conditions. The subjects were compared on learning gains, satisfaction, learning rate, and how much of their error rate during tutoring is explained by the number of opportunities to practice. We found that subjects learn KC 32 (independence of x and y motion) better when seeing only illustrations and that subjects seeing only illustrations prefer the graphics more than when seeing graphs or multiple representations.

However, when looking at interaction effects with incoming knowledge and gender, we found that, for some students, multiple representations are better. Males and subjects with low incoming knowledge might benefit from multiple representations. For males and subjects with no calculus experience, seeing multiple representations produced larger learning gains than seeing a single representation on KC 25 (horizontal component of velocity in projectile motion is constant). Subjects with no calculus experience also learned faster and had more of their decrease in error rate described by an increase in practice when seeing multiple representations. Subjects with no high school physics preferred the post-problem discussions more when seeing multiple representations.

This study was also designed to address the question of whether an adaptive scheduling of representations is more beneficial to learners than a fixed, alternating scheduling. In answering this question, I compared subjects in the adaptive condition to subjects in the alternating condition. Unfortunately, since the number of switches between representations is strongly dependent on the condition (adaptive subjects saw between one and four, alternating subjects saw twelve), any findings can be the result of the frequency of switching and not the method of scheduling. When comparing conditions on learning gains, satisfaction, learning rate, and how much of their error rate during tutoring is explained by the number of opportunities to practice, we find no significant differences. This could suggest that adaptive scheduling is no better than alternating. However, it may also suggest that the frequency of switching is not important, which would contradict other research [89, 91], just that subjects see multiple representations. Additionally, it may suggest that the frequency of switching can be reduced if done according to an adaptive policy.

Examining interaction effects with pretest score, high school physics, and calculus finds cases where the adaptive policy is better. For low pretesters, subjects learn more on KC 25 (horizontal component of velocity in projectile motion) and like the overall tutoring system more with the adaptive schedule than the alternating schedule. However, high pretesters learn faster when receiving an adaptive schedule and low pretesters learn faster when receiving an alternating schedule. Subjects with high school physics learn more on KC 15 and KC 23 (both involving acceleration) with an adaptive schedule over an alternating schedule. Subjects with calculus experience learn quicker and have more of their learning explained by practice opportunities when seeing an adaptive scheduling.

Generally, subjects with high incoming knowledge perform better when receiving an adaptive schedule, whereas those with low incoming knowledge perform better with an alternating schedule. One exception is that low pretesters perform better on KC 25 and like the overall system more when receiving an adaptive scheduling. This seeming contradiction could be explained by different knowledge components benefit differently from adaptive versus alternating policies (or the frequency of switching). However, more work is needed to investigate this.

Finally, this study was also an extrinsic evaluation of the adaptive policy learned in the previous chapter. Since the adaptive condition does not consistently perform better, I conclude that the adaptive policy needs improvement. The policy was learned from data where subjects never switched representations during tutoring. However, it was used to determine when to switch representations. Thus, I believe that the adaptive policy was limited in what tutoring contextual information it could consult.

To test this belief, I created nine new features to capture tutoring context information regarding the switching of representations. Using both the old features and these nine new features, a new policy was learned from the data collected in this study. The new policy used seven of the nine new features and almost half of the policy rules involved these new features. It performed better on an intrinsic evaluation than two baselines, suggesting that the representational information captured by these features is important for an adaptive policy. Rules from the policy also appear consistent with the literature. Subjects who can handle graphs or who are doing well should see graphs [57, 70, 69]. There should also be a

general trend over time from showing illustrations to showing graphs [57, 83]. Now that it has been shown that these features are important, in future work I propose searching for a better model.

One of the two baselines was based off of the best model learned in Chapter 5: PreScore* Gender model. I also used the data to relearn the weights for the model. This model performed significantly better than a single-graphic baseline, suggesting that the modeling approach outlined in Chapter 5 may transfer to new situations, but the weights may need to be relearned.

6.3.1 Lessons Learned

In augmenting the Rimac tutoring system with visual representations, I learned lessons that may be helpful in future research along these lines.

Two common complaints about the tutoring environment involved the displays. One complaint involved the refresh rate of the monitors, which were old cathode ray tube (CRT) monitors. These monitors had refresh rates of 48 Hz. Some subjects commented that they could see slight flickering on the screen, which they found distracting. A few students also commented towards the end of the study that their eyes were tired, which may also be due to the flickering. It might be good for future studies to use monitors with higher refresh rates to possibly reduce distractions and eye fatigue.

The second common complaint about the displays was that the text in the images was too small. The font family and size (Arial 9-point font) were chosen because they were the defaults in Gnuplot, the graphing software used to generate the graphs. For the illustrations to be consistent, text in the illustrations were also 9-point Arial. Some students remarked that the text in the graphics was too small to read easily. In the future, it might be good to use a larger font size in graphics.

Students who saw illustrations occasionally commented that the illustrations were nicer than what they've seen in some science classess. This is in contrast to comments from the first study, where they found the illustrations too abstract. The addition of extraneous features such as a sun in the background, a blue sky, and green grass seemed to help to make

the illustrations more concrete.

Augmenting Rimac with graphics was a fairly smooth process. The underlying TuTalk authoring language already supported embedding graphics into dialogues. The hard part was that the Rimac client did not support graphics. This led me to use an older TuTalk client, Cordillera. The advantage to using the Rimac client was that the Andes physics homework helper could be embedded within it, giving a consistent interface during tutoring. Cordillera did not support embedding Andes within it, so a separate web browser was used to display Andes, with only part of Cordillera visible for the walkthrough dialogues; the rest of Cordillera was hidden off-screen. When it came time for the reflection dialogues, all of Cordillera would be made visible so the subject could see both the dialogue and the graphics, which were in a separate pane of the interface.

Table 6.30: Summary of findings from the chapter.

Single vs. Multiple Representations		
Interaction	Measure	Finding
None	KC32	illustrations are better than multiple representations
None	PS Graphics	multiple representations and illustrations are better than graphs
None	PPD Graphics	multiple representations are better than graphs
None	Graphics	multiple representations and illustrations are better than graphs
HS Physics	PPD	Subjects with high school physics do better with single representations Subjects without high school physics do better with multiple representations
Calculus	KC25 Learning Rate Adj. r^2	Subjects with calculus experience do better with single representations Subjects without calculus experience do better with multiple representations
Gender	KC25	Females do better with single representations Males do better with multiple representations
Alternating vs. Adaptive		
Interaction	Measure	Finding
None	KC24	Frequency of switching adversely affects learning
Pretest	KC25 Overall Satisfaction	Low pretesters do better with adaptive scheduling High pretesters do better with alternating scheduling
Pretest	Learning Rate	Low pretesters do better with alternating scheduling High pretesters do better with adaptive scheduling
High School Physics	KC15 KC23	Subjects with high school physics do better with adaptive scheduling Subjects without high school physics do better with alternating scheduling
Calculus	Learning Rate Adj. r^2	Subjects with calculus experience do better with adaptive scheduling Subjects without calculus experience do better with alternating scheduling

7.0 CONTRIBUTIONS AND FUTURE WORK

Physics education research and cognitive science research both suggest that different graphical representations provide differing benefits to learners [81, 69, 84, 39, 70], suggesting a need to adapt representations to learners in physics tutoring systems. However, physics tutoring systems tend to rely on a single graphical representation (e.g. [109, 3, 7, 58, 62]). Tutoring systems that adapt to students do not use multiple representations as part of the adaptation (e.g. [6, 33, 52, 40, 50]), and tutoring systems that use multiple representations do not adapt those representations and their presentation to individual students (e.g. [57, 83, 91]). This research is positioned at the intersection of two areas of research: using multiple graphical representations in tutoring and tutoring systems that adapt to learners.

In the research presented, I explored the importance of adapting the scheduling of multiple graphical representations to learners and tutoring contexts in the context of a dialogue-based conceptual physics tutor. The work was designed to test three overarching hypotheses. H1: determine whether the best graphical representation for physics concepts is related to properties of the learner and the tutoring context. H2: examine whether it is possible to build models that predict the best representation using those properties. H3: show that using an adaptive scheduling of representations based on a model learned while investigating H2 will produce greater learning gains than a non-adaptive scheduling. In working on H2, I developed a modeling algorithm that augments multiple linear regression to syntactic constraints. In testing H3, I also tested a fourth hypothesis, H4: show that multiple graphical representations are more beneficial for learning than a single graphical representation.

7.1 H1: BEST REPRESENTATION DEPENDENT ON LEARNER AND TUTORING CONTEXT

Within the artificial intelligence in education community, it was not known whether differences existed between illustrations and graphs for conceptual physics learning. Prior research in the physics education community had shown that illustrations and graphs have different learning benefits for different students [81, 69, 84], but these focused on paper-and-pencil learning and without a natural-language-based tutor. The artificial intelligence in education community has made use of illustrations and graphs but without considering the differing benefits of each.

In Chapter 4, data was collected from a study in which subjects saw either illustrations during tutoring or graphs. During tutoring, subjects solved problems in the Andes intelligent homework helper [109], then discussed concepts related to the problems in an augmented version of the Rimac reflection dialogue tutor [65].

From this data, interaction effects were identified for learning gains between representation seen and student features, confirming H1. Low pretesters seeing illustrations during tutoring score higher than low pretesters seeing graphs; this result is consistent with findings from cognitive science [70, 57]. However, for high pretesters, there is no consistently good representation, contrary to other research which suggested that non-novice learners would perform better with graphs [57]. Instead, it may suggest that switching representations will be beneficial for them, which is in partial agreement with related work suggesting that switching representations are beneficial for all learners [91].

In using the data to build models from which adaptive policies could be created, we find tutoring contexts when one representation may be more beneficial than the other. We again see that low pretesters benefit most from illustrations. For high pretesters, we find further evidence that switching representations is beneficial for them. However, instead of a fixed switching schedule, as suggested in [91], we find particular contexts when one representation is more beneficial than the other. Illustrations are better if there are many correct responses in the tutoring session or if it is later in tutoring and graphs are better if there are fewer correct responses. These findings are in opposition to prior work, which

suggests showing abstract representations such as graphs later in tutoring [57] or when the student is doing well [104, 69] and showing illustrations when the student is performing poorly [81, 57]. These opposite findings may be due to differences in domains (e.g. [57, 104]) or different situations (e.g. problem solving [69]). However, it may also be that there are additional factors influencing these results or that the model identifies tutoring contexts in which illustrations or graphs present a challenge to the learners, which with the help of the natural language tutor learners are able to overcome. More research is needed to determine the cause of the different results found in this body of research and prior research.

We also found that subjects with low spatial reasoning perform significantly better when viewing graphs during tutoring than illustrations. While this supports H1, it is contradictory to cognitive science literature which suggests that low spatial reasoning subjects would perform poorly when given graphs to learn from [70]. Although the cognitive science literature suggested high spatial reasoning subjects would show no difference in learning when shown illustrations versus graphs [70], we found that they performed better with illustrations. One possible explanation is that students who saw illustrations during tutoring needed to work harder to learn the concepts in a context-free way. Graphs provide a context-free representation of the concepts [74, 104] and so students in this condition may not have needed to work as hard to learn the concepts in a context-free way. Another possibility is that the test used for determining spatial reasoning ability – the Paper Folding Test, a standard test for measuring spatial reasoning ability [47, 20] – may not have accurately measured the spatial reasoning skills used in this study. Further research is needed to determine which spatial reasoning test best measures the spatial reasoning skills used in interpreting graphs for physics.

We find additional learner and tutoring context factors influencing the best representation from the adaptivity model for pretest and gender from Section 5.3.2. Low pretesting females may benefit from seeing graphs when performing well, whereas low pretesting males may benefit from illustrations. Female high pretesters may benefit from graphs early in tutoring, whereas female low pretesters may benefit from illustrations; this is consistent with the literature on when to show illustrations and graphs [57, 83].

However, the low adjusted r^2 values in the models learned in Chapter 5 suggest that

not all learning is explained by the features used in the models. While this may be due to limitations of the modeling technique, it may also be due to unidentified features that are important to learning. Identifying these new features that explain the learning is an avenue for future work. Section 6.2.3 began to identify some features, noting that how subjects performed previously with a given representation can have a significant effect on which representation to show them in the future. The findings in Section 6.2.2.3 that certain concepts prefer different schedulings may suggest that the concepts being tutored are better-learned using one representation or scheduling of representations than another. More research is needed to investigate the effect concepts have on the representation to display. Additionally, more research is needed to identify additional features.

7.2 H2: BUILDING MODELS TO PREDICT REPRESENTATION

Since adapting graphical representations to learners and tutoring contexts had not been done before, one contribution of this work is showing that models can be built that predict the better representation for a particular learner and tutoring context. To build the models, a new modeling algorithm was developed. The features used in the modeling had been found useful in other student modeling tasks [42, 11, 25] and have been found useful in cognitive science research on visual representations [84, 70]. I show that many of these features are useful in predicting when to show illustrations and when to show graphs.

In building the models, there was no gold standard; we had only which representation a subject saw and how much they learned. Thus, to build a model, I developed an algorithm which modifies stepwise linear regression according to syntactic constraints, similar to other student modeling approaches [22, 86, 24]. The models learned predict post-test score given pretest score and a collection of terms representing context information. The terms are a product of binary features describing learner and tutoring context information; one feature in each term indicates whether a graph or illustration was shown. Each term can be converted into a rule indicating which contexts illustrations or graphs are beneficial. After stepwise regression selects the terms that explain the change in post-test score, the modeling

algorithm must then identify problematic pairs of terms (e.g. contradictory rules, such as show illustrations for low pretesters and show graphs for low pretesters) and remove the term with the lower absolute value of the coefficient. Linear regression is run on the remaining terms to determine the final coefficient values. The terms are then ranked according to the absolute value of the coefficient (higher value indicates more important) and converted into adaptive policy rules.

For an intrinsic evaluation, the models learned were compared to a single-representation model baseline on adjusted r^2 values. The learned model using only one context feature (plus a feature indicating whether to show illustrations or graphs) in each term performed no different than the baseline. However, adding additional context in each term improved performance. I found that including a feature indicating pretest score and a feature indicating gender produced the model with the highest observed adjusted r^2 value (the PreScore*Gender model). This is in agreement with the cognitive science literature, which suggested that gender and knowledge were important to consider when deciding which representation to use [104, 69, 84, 70].

To evaluate whether the model can transfer to new situations, the data from the study described in Chapter 6, was used to compare the adaptivity based on pretest and gender to a non-adaptive model baseline. The terms in the adaptive model model were kept, but the weights were relearned using the new data. It performed significantly better than the baseline, showing that the modeling approach outlined in Chapter 5 may transfer to new situations, although coefficients may need to be relearned.

An extrinsic evaluation of one of the models learned occurred in the user study described in Chapter 6. Here, the model was compared to a non-adaptive model that just alternated between the representations. Since the adaptive model does not consistently perform better, this suggests that either the third overarching hypothesis is wrong or that the policy needs improvement. The policy was learned from data where subjects never switched representations during tutoring. However, it was used to determine when to switch representations. Thus, I believe that the adaptive policy was limited in what tutoring context information it could consult. The small pilot experiment in Section 6.2.3 supports this.

In future work, improvements to this modeling technique should be explored. One avenue

is the automatic selection of “important” features to use when adding factors to a term to increase contextual description. In this work, I consulted the literature to pick these features, but an automatic method may identify unexpected but important features. More work should also be done to determine how best to identify and handle problematic rules. Preliminary work was done in this thesis, but more can be done, such as whether running multiple linear regression is the best last step or whether stepwise regression should be run again. Another avenue for improving the algorithm is to determine how best to handle non-adaptive rule pairs. In this work, I removed the rule with a lower *avc*, but another possibility is to merge the two rules, removing the factor that is opposite in the two rules. Finally, this modeling algorithm should be compared to other algorithms such as reinforcement learning, which is another technique used to develop models when the gold standard is not available.

7.3 H3 AND H4: BEST SCHEDULING OF REPRESENTATIONS

Within the artificial intelligence in education community, it is not known whether adapting graphical representations to learners and tutoring contexts is more beneficial for learning over alternating graphical representations. Evidence from the cognitive science and physics education literature suggest that adapting would better [81, 69, 84, 39, 70], as does the evidence supporting hypotheses H1 and H2, but no tutoring study had yet investigated it. Additionally, it is not known whether multiple graphical representations are more beneficial for learning than a single graphical representation for conceptual physics tutoring and when accompanied with a natural language representation. Research within the artificial intelligence in education community on other subjects found that multiple graphical representations were better than a single graphical representation [57, 83, 91], but no such research had yet been performed for physics and with the graphical representations accompanied by natural language dialogues. In Chapter 6, I investigate both hypotheses.

In examining H3, subjects receiving an adaptive scheduling of representations were compared to subjects receiving an alternating scheduling. We find no significant differences between subjects. This could suggest that adaptive scheduling is no better than alternat-

ing. However, there was a significant difference in the frequency of switching representations between the two conditions. Therefore, since no significant differences were found between subjects in the two conditions, it may also suggest that the frequency of switching is not important, which would contradict other research [89, 91].

However, some subjects seem to benefit from adaptive scheduling. Generally, subjects with high incoming knowledge perform better when receiving an adaptive schedule, whereas those with low incoming knowledge perform better with an alternating schedule. One exception is that low pretesters perform better on concepts involving projectile motion and like the overall system more when receiving an adaptive scheduling. This seeming contradiction could be explained by different knowledge components benefit differently from adaptive versus alternating policies (or the frequency of switching). A future study is needed to determine whether certain concepts benefit from one representation over another or one scheduling of representations over another.

While adaptive scheduling has benefits for some subjects, it is not clear that adaptive scheduling is better than alternating scheduling. This may be due to the adaptive policy used for the adaptive scheduling. The policy was trained from data where subjects only ever saw one representation, but was then used to predict which representation a subject should see. The target application did not match the source application of the training data. Thus, the adaptive policy may be failing to capture aspects of the tutoring context that are important for determining which representation to show; the small pilot experiment in Section 6.2.3 supports this. A future study should be performed that controls the frequency of switching for both an alternating scheduling and an adaptive scheduling. A better policy should also be learned from the data collected in Chapter 6. Alternatively, a new study could be performed that randomly decides which representation to show. This random ordering of representations for each subject would yield a wide variety of representation sequences and tutoring contexts for those representations. This would increase the search space that the modeling algorithm has to identify important contexts for each representation, potentially leading to a better adaptive model learned.

In examining H4, subjects seeing only a single representation were compared to subjects seeing multiple representations (scheduled either adaptively or alternating). Subjects seeing

only illustrations learn the independence of x and y motion better and prefer the graphics more than subjects seeing multiple representations and subjects seeing only graphs. While this result seems counter to the literature on multiple graphical representations, it may be that the subjects can all be considered novices since none of them have had college-level physics. Some research suggests that novices benefit more from illustrations over graphs [70].

However, some students do perform better with multiple representations. Males and subjects with no calculus experience had larger learning gains on questions related to projectile motion when seeing multiple representations. Subjects with no calculus experience also learned faster and had more of their decrease in error rate during tutoring described by an increase in practice when seeing multiple representations. Subjects with no high school physics preferred the post-problem discussions more when seeing multiple representations.

While this study found that subjects generally perform better with a single representation, the multiple representations were scheduled to be either alternating or adaptive. One option for scheduling multiple representations that was not evaluated was concreteness fading – transitioning from a concrete representation (e.g. illustrations) to an abstract representation (e.g. graphs). In a future study for natural-language based conceptual physics tutoring, the single-representation condition should be compared to concreteness-fading scheduling of multiple representations. Additionally, this study compared single and multiple graphical representations within the context of a typed dialogue tutoring system. It is possible that the added dimension of a typed dialogue affects the benefits of multiple graphical representations. A future study should investigate this, perhaps by comparing a dialogue-only condition, a multiple-graphical-representations-only condition, and a condition with both dialogue and multiple graphical representations.

7.4 GENERALIZING RESULTS

While this work was consistent with some prior work, it also had findings that contradicted other prior work. Thus, it is important to discuss how this work differs from prior work,

to help explain why differences were found with prior work and to offer suggestions on the generalizability of the findings from this work. The results presented in this dissertation were found within the context of a natural language-based conceptual physics tutor, with communication between subject and computer tutor occurring through text. Related work cited in this dissertation differ on various points, such as involving fractions instead of physics [91], or on physics test rather than tutoring [70]

Where findings presented here agree with this work suggest areas where the findings may generalize. Where the findings disagree suggest there may be specific features of the domain, tutoring environment, or context where the findings are specific to the domain, environment, or context. Future work could investigate what it is that causes those differences to lead to different findings.

In developing the models, we find that gender, pretest score, and correctness during tutoring are important features for predicting which representation to use. These features are also important for determining student performance on solving kinematics [69, 70], work-energy [104], and electrical circuit [84] physics problems; and for a conceptual physics work-energy tutor [25]. The diversity of physics topics and problem types (i.e. problem solving and conceptual) suggests that these features are helpful across physics topics and problem types. These features are also important for determining student affect in a spoken dialogue computer tutor for computer literacy [42]; determining hint interactivity in a biology tutor [11]; and in determining pedagogical strategies for a conceptual physics tutor [25]. This suggests that the features are generally helpful for many different student modeling tasks. Combined, these features are likely helpful for many modeling tasks.

We found that low pretesters perform better when seeing illustrations. This is consistent with research on learning scientific principles [57], physics problem solving [70], and mathematical knowledge [83]. Given the diversity of subjects, the finding likely generalizes to math and science topics in general.

For high pretesters, there does not appear to be a consistently good representation, contrary to prior work [57]. Contrary to other work, we find that the frequency of switching is not important [89, 91]. Instead, we find particular contexts when one representation is better than another. These findings are in opposition to prior work, which suggests showing

abstract representations such as graphs later in tutoring [57] or when the student is doing well [104, 69] and showing illustrations when the student is performing poorly [81, 57]. These opposite findings may be due to differences in domains (e.g. [57, 104]) or different situations (e.g. problem solving [69]). More research is suggested to determine the cause of the different results found in this dissertation and in prior research.

APPENDIX A

BACKGROUND SURVEY

- Age
- Year in School
 - Freshman
 - Sophomore
 - Junior
 - Senior
 - Other (please specify)
- Gender
 - Female
 - Male
- Academic Major
- Math SAT
- Reading SAT
- Writing SAT
- Did you take physics in high school?
 - No
 - Yes

- Advanced Placement
- Have you taken (or are you taking) Calculus? If you’ve taken calculus in high school, choose “Yes”.
 - No
 - Yes
- Do you plan to take physics some time during your college years?
 - No
 - Maybe
 - Yes

APPENDIX B

TEXTBOOK

B.1 INTRODUCTION

This brief textbook was designed to provide you with sufficient background knowledge of physics to be able to learn something from this study. It assumes no prior college-level physics knowledge.

A Note About Notation: For consistency's sake, this textbook shows equations and numbers that use superscripts and subscripts the same way you would type them in this study.

For example, to enter an equation with exponents, you would indicate the superscripted exponents using the carat character \wedge (*shift-6*):

$$a^2 + b^2 = c^2 \quad \text{---} \quad a^2 + b^2 = c^2$$

Similarly, to enter an equation with subscripts, you would indicate the subscripts using adjacent [lower-case] letters:

$$v_f = v_i + at \quad \text{---} \quad v_f = v_i + a^*t$$

To enter an expression involving a square root, use **sqrt**:

$$c = \sqrt{a^2 + b^2} \quad \text{---} \quad c = \text{sqrt}(a^2 + b^2)$$

$$v_0 = \sqrt{KE_0/(1/2m)} \quad \text{---} \quad v_0 = \text{sqrt}[KE_0 / (0.5*m)]$$

B.2 DISPLACEMENT

Displacement is a vector quantity representing the change in position of an object. Its standard unit of measure is the meter (m) and is often represented by a variable name beginning with d.

It is not the same as the scalar measure of distance traveled by a moving object. For example, a swimmer who swims in a straight line from one end of a 100-m pool to the other end and back swims a total distance of 200 m, but her displacement during that same time interval is 0 m because she returns to her original position.

If a football player runs straight down the sideline from one end zone to the other, he runs a total distance of 100 yds, and his displacement is 100 yds in the direction of the far end zone. However, if he starts and ends at the same two points on the sideline as before, but runs in a zig-zag or haphazard pattern all over the field instead of a straight line, his displacement is still 100 yds downfield but he will have run a distance greater than 100 yds.

B.3 VELOCITY

Velocity is a vector quantity representing the rate of change in position of an object per unit time. Its standard unit of measure is meters per second (m/s) and is often represented by a variable name beginning with v.

The *average* velocity of a moving object is its displacement during a time interval divided by the duration of the time interval. However, velocity in this study will refer to the *instantaneous* velocity of an object, which is its displacement per unit time at a given point in time (or, over an infinitesimal time interval).

For example, let's say you throw a ball straight up in the air with an initial velocity of 5 m/s upwards. A short time later, its velocity will be less than 5 m/s upwards because gravity will slow it down. At the point in time where the ball ceases to move upwards but before it begins to fall back down, it has a velocity of 0 m/s. It will eventually land back in your hands with a downward velocity of some magnitude.

B.4 ACCELERATION

Acceleration is a vector quantity representing the rate of change in velocity of an object per unit time. Its standard unit of measure is meters per second per second, or meters per second squared (m/s^2) and is often represented by a variable name beginning with a.

The *average* acceleration of a moving object is its change in velocity during a time interval divided by the duration of the time interval. However, acceleration in this study will refer to the *instantaneous* acceleration of an object, which is its change in velocity per unit time at a given point in time (or, over an infinitesimal time interval).

As with other vector variables, be careful to account for direction when dealing with acceleration. For example, suppose the driver of a car moving east at 60 mph applies the brakes. The car's velocity vector will continue to be eastward until the car stops, but the braking acceleration vector will be to the *west*, in the direction opposite the car's displacement. In other words, because it's slowing down, the acceleration must be in the opposite direction of velocity. Had they been in the same direction, then the car would be speeding up.

Although some vectors change in magnitude or direction as an object moves (such as the velocity on this and the previous page), most of the accelerations you will encounter in this study will be *constant*.

B.5 GRAVITATIONAL ACCELERATION

One such constant acceleration used in this study is the acceleration due to gravity of a moving object near the surface of the Earth. It can be shown experimentally that all objects near the Earth's surface have exactly the same downward acceleration, whenever the effects of air resistance can be eliminated or otherwise ignored.

Unlike many other accelerations, this special constant is *not* represented by a variable name beginning with a. Its magnitude is represented by the scalar variable g , and its direction is always straight down (toward the center of the Earth): $g = 9.8 \text{ m/s}^2$ downward

B.6 RELATIONSHIPS BETWEEN QUANTITIES

The physical quantities of displacement, velocity, acceleration, and time are related through a few equations. You do not need to memorize these equations; you will have access to them when you need them.

Table B1: Equations and descriptions for one dimensional motion.

Equations	Description
$d = d_0 + v_0 * t + 1/2 * a * t^2$	The displacement (d) at time (t) is dependent on the initial displacement (d0), the initial velocity (v0), and the acceleration (a). Usually initial velocity (v0) is zero.
$d = d_0 + v_0 * t$	If there is no acceleration, the displacement (d) at time (t) is dependent on just the initial displacement (d0) and the initial velocity (v0).
$v = v_0 + a * t$	The velocity (v) at time (t) is dependent on the initial velocity (v0) and the acceleration (a).
$v = (d_1 - d_0) / t$	The average velocity (v) depends on the change in displacement (d1 - d0) over the time it took for that change to take place.
$v^2 = v_0^2 + 2 * a * d$	The velocity (v) depends on the initial velocity (v0), acceleration (a), and displacement (d).
$a = (v_1 - v_0) / t$	The average acceleration (a) depends on the change in velocity (v1 - v0) over the time it took for that change to take place.

Note: For this study, if the movement is in the vertical (y) direction, then the acceleration

(a) can be replaced with gravity (g) in the equations above.

B.7 TWO DIMENSIONAL MOTION

Since displacement, velocity, and acceleration are vector quantities, they have a direction. For the problems you will be seeing in this study, there will be two main dimensions:

- horizontal (x-direction): motion left and right, such as a child running down the street
- vertical (y-direction): motion up and down, such as a child tossing a ball into the air

Horizontal motion and vertical motion are independent of each other. For example, when a child tosses a ball into the air, its height is not affected by whether or not the child is running. Similarly, whether or not a child tosses a ball into the air does not affect how far the child runs.

Since gravity is a downward acceleration, it only acts in the vertical direction (y-direction) not in the horizontal (x-direction).

Examples of two-dimensional motion include dropping an object, tossing an object up and letting it fall, or kicking a football down the field.

In each of these cases, the objects are being pulled down by gravity, even if they were initially going up (as in the child tossing the ball or a ball being kicked). In the third example, of the football being kicked, it was given an initial velocity in both the x-direction (horizontal) and y-direction (vertical), but gravity only acted in the y-direction.

Since the x-direction and y-direction act independently of each other, it is possible to separate the x-component and y-component of the physics vectors and talk about them separately. To refer to the:

- x-component, we use the subscript x
- y-component, we use the subscript y

So, to refer to the x-component of displacement, it would be \mathbf{d}_x . The y-component of initial velocity would be $\mathbf{v0}_y$.

This two-dimensional motion is called projectile motion because there is a projectile. In the case of a ball being tossed to someone else, the projectile would be the ball. On earth, the ball will initially go upward, but will eventually arc downward. This is because gravity is pulling it down. Had the ball been tossed in space, where gravity doesn't exist, it will continue going in the direction it was tossed – it won't arc downward!.

B.8 RELATIONSHIPS BETWEEN QUANTITIES (2D)

Here we see how the equations change when considering two-dimensional motion. You do not need to memorize these equations; you will have access to them when you need them.

Notice:

1. We're repeating most of the same equations from before. There are fewer equations in the Y-Direction because there will always be an acceleration, so the equations with no acceleration aren't in the Y-Direction table.
2. In the y-direction, acceleration is due to gravity for this study.
3. Time does not have an x-component or a y-component!

X-Direction (Horizontal motion)

Table B2: Equations and descriptions for the horizontal component of projectile motion.

Equation	Description
$d_x = v_{0_x} * t + 1/2 * a_x * t^2$	The displacement (d_x) at time (t) is dependent on the initial velocity (v_{0_x}) and the acceleration (a_x).
$d_x = v_{0_x} * t$	If there is no acceleration, the displacement (d_x) at time (t) is dependent on just the initial velocity (v_{0_x}).
$v_x = v_{0_x} + a_x * t$	The velocity (v_x) at time (t) is dependent on the initial velocity (v_{0_x}) and the acceleration (a_x).
$v_x = (d1_x - d0_x) / t$	The average velocity (v_x) depends on the change in displacement ($d1_x - d0_x$) over the time it took for that change to take place.
$v_x^2 = v_{0_x}^2 + 2 * a_x * d_x$	The velocity (v_x) depends on the initial velocity (v_{0_x}), acceleration (a_x), and displacement (d_x).
$a_x = (v1_x - v0_x) / t$	The average acceleration (a_x) depends on the change in velocity ($v1_x - v0_x$) over the time it took for that change to take place.

Y-Direction (Vertical motion)

Table B3: Equations and descriptions for the vertical component of projectile motion.

Equation	Description
$d_y = v_{0_y} * t + 1/2 * g * t^2$	The displacement (d_y) at time (t) is dependent on the initial velocity (v_{0_y}) and gravity (g).
$v_y = v_{0_y} + g * t$	The velocity (v_y) at time (t) is dependent on the initial velocity (v_{0_y}) and gravity (g).
$v_y = (d_{1_y} - d_{0_y}) / t$	The average velocity (v_y) depends on the change in displacement ($d_{1_y} - d_{0_y}$) over the time it took for that change to take place.
$v_y^2 = v_{0_y}^2 + 2 * g * d_y$	The velocity (v_y) depends on the initial velocity (v_{0_y}), acceleration (g), and displacement (d_y).

APPENDIX C

SATISFACTION SURVEY QUESTIONS

This appendix shows the satisfaction survey questions used.

Problem-Solving Graphics

1. I found the graphics easy to understand.
2. I thought that the graphics helped me understand the problem better than not having them.
3. When solving the physics problem, I referred to the graphic displayed to help me solve it.
4. The graphics displayed are the kind I like to see when solving a physics (or math or science) problem.

Post-Problem Discussion

1. I thought working with the post-problem tutor was fun.
2. I felt in control of my conversations with the post-problem tutor.
3. I felt like the post-problem tutor adapted to my responses.
4. I was able to recover easily from errors in the post-problem tutor.
5. I found the information provided by the post-problem tutor to be clear.
6. I thought that the post-problem discussion was clear and natural.
7. I found it easy to learn from the post-problem tutor.

8. My conversation with the post-problem tutor quickly led to me having a deeper understanding of the material.

Post-Problem Discussion Graphics

1. I found the graphics easy to understand.
2. I liked that the graphics displayed during the discussion were the same as the graphics during problem-solving.
3. I feel like the graphics helped me understand the concepts better than not having them.
4. I thought that the graphics were well-integrated into the dialogues.
5. When learning physics (or math or science), I prefer seeing the kinds of graphics that I saw.
6. If I had trouble understanding the graphics, the tutor explained them clearly.

Overall System

1. I am satisfied with my experience learning physics from this system.
2. I thought that working in this learning environment was just like working one-on-one with a human tutor.
3. I would use this system again to continue learning about physics.
4. I would like to be able to use a system like this to learn about other topics.
5. I think this was a good way to learn about physics.

Open-Ended Questions

1. What did you like the *most* about the graphics and post-problem discussion?
2. What did you like the *least* about the graphics and post-problem discussion?
3. What would you like to see changed or improved about the graphics and post-problem discussion?
4. Is there anything else you would like to say?

APPENDIX D

TEST QUESTIONS FOR THE FIRST STUDY

This appendix shows the two tests for the first study, described in Chapter 4. In italics and parentheses following each question is the tag for the type of question.

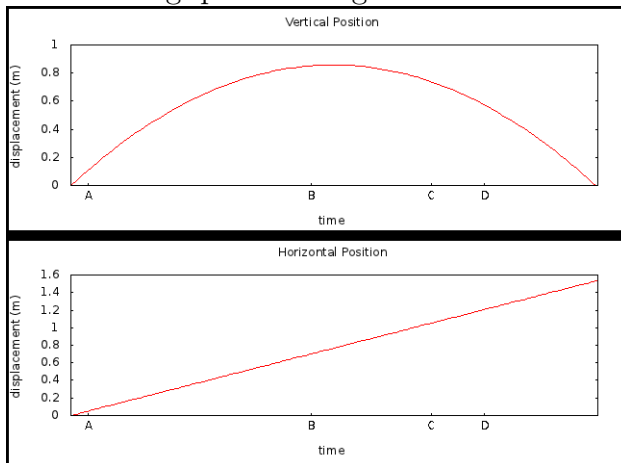
D.1 TEST A

This section shows the test questions for test A.

1. An orange rolls off the edge of a table and lands on the floor. The table is 1.50 m high and the orange lands 0.6 m from the base of the table.
 - a. Which of the equations below would you use to determine how long it took for the orange to fall? (*numeric*)
 - $d_y = v_{0y} * t + 1/2 * g * t^2$ (in Andes: $d_y = v_{0_y} * t + 1/2 * g * t^2$)
 - $v_y^2 = v_{0y}^2 + 2 * g * d_y$ (in Andes: $v_y^2 = v_{0_y}^2 + 2 * g * d_y$)
 - $v_y = v_{0y} + g * t$ (in Andes: $v_y = v_{0_y} + g * t$)
 - $g = (v_y - v_{0y})/t$ (in Andes: $g = (v_y - v_{0_y}) / t$)
 - b. Which number would you use for v_{0y} ? (*numeric*) · 0 m/s
 - 1.50 m
 - 0.6 m
 - 9.8 m/s²

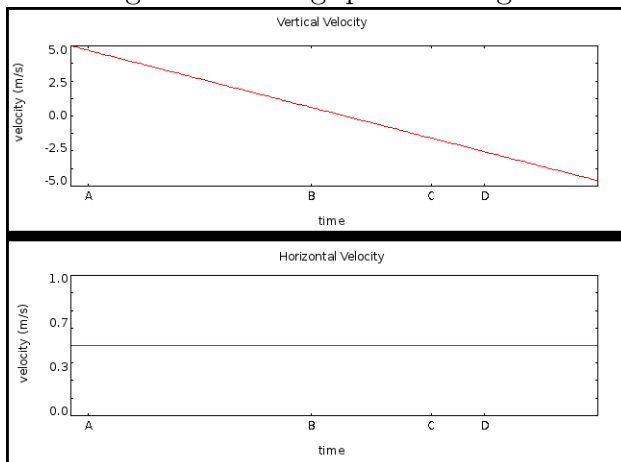
- That value is not known for this problem.
- c. Assuming you have solved for time already (in part a), which of the equations below would you use to determine how fast the orange was rolling as it rolled off the edge? (*numeric*)
 - $v_y = v_{0y} + g * t$ (in Andes: $v_y = v0_y + g * t$)
 - $v_x^2 = v_{0x}^2 + 2 * a_x * d_x$ (in Andes: $v_x^2 = v0_x^2 + 2 * a_x * d_x$)
 - $d_y = v_{0y} * t + 1/2 * g * t^2$ (in Andes: $d_y = v0_y * t + 1/2 * g * t^2$)
 - $d_x = v_{0x} * t$ (in Andes: $d_x = v0_x * t$)
- d. What physics quantity represents change in velocity over time? (*conceptual*)
 - displacement
 - acceleration
 - velocity
 - None of the above

2. A child tosses a ball up and forward. The two graphs shown below show the vertical and horizontal positions of the ball over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.



- a. Where is the magnitude of the vertical velocity largest? (*graphs*)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal

- b. Where is the magnitude of the horizontal velocity largest? (*graphs*)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal
- c. Where is the magnitude of vertical velocity smallest? (*graphs*)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal
3. A child tosses a ball up and to the right (just like on the previous page). Now, the graphs below show the vertical and horizontal **velocities** over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.



- a. Between time A and time B, what is the direction of vertical **velocity**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving vertically
- b. Between time A and time B, what is the direction of vertical **acceleration**? (*graphs*)

- up
 - down
 - left
 - right
 - the ball isn't moving vertically
- c. Between time C and time D, what is the direction of vertical **velocity**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving vertically
- d. Between time C and time D, what is the direction of vertical **acceleration**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving vertically
- e. Between time A and time B, what is the direction of horizontal **velocity**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving horizontally
- f. Between time A and time B, what is the direction of horizontal **acceleration**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving horizontally

4. A woman tosses her keys straight up into the air.
- On the way up**, after it leaves her hand, what is the direction of velocity? (*conceptual*)
 - Up
 - Down
 - Left
 - Right
 - There is not enough information to know.
 - On the way up**, after it leaves her hand, what is the direction of acceleration? (*conceptual*)
 - Up
 - Down
 - Left
 - Right
 - There is not enough information to know.
 - On the way back down**, what is the direction of velocity? (*conceptual*)
 - Up
 - Down
 - Left
 - Right
 - There is not enough information to know.
 - On the way back down**, what is the direction of acceleration? (*conceptual*)
 - Up
 - Down
 - Left
 - Right
 - There is not enough information to know.
5. Suppose that a freely falling object were somehow equipped with a speedometer. By how much would its speed reading increase with each second of fall? Ignore air resistance. (*numeric*)

- 19.6 m/s
- 9.8 m/s
- 4.9 m/s
- 0.0 m/s, the speed is constant
- None of the above.

6. A jogger is jogging west towards a water fountain.

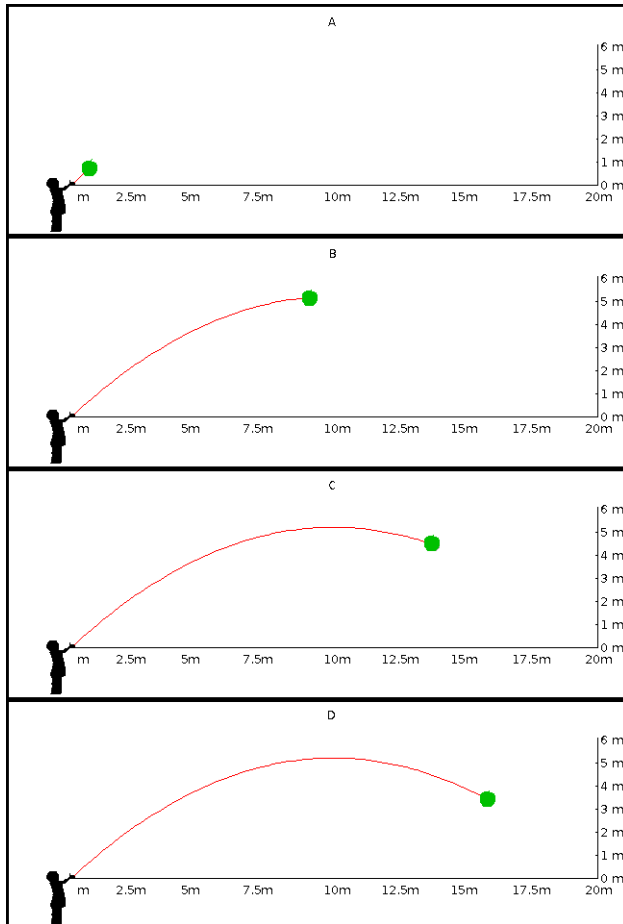
a. As she approaches the fountain, she slows to a stop. What is the direction of acceleration? (*conceptual*)

- East
- West
- There is no acceleration
- None of the above

b. After getting a drink of water, she continues jogging west. What is her direction of acceleration as she returns to her normal jogging speed? (*conceptual*)

- East
- West
- There is no acceleration
- None of the above

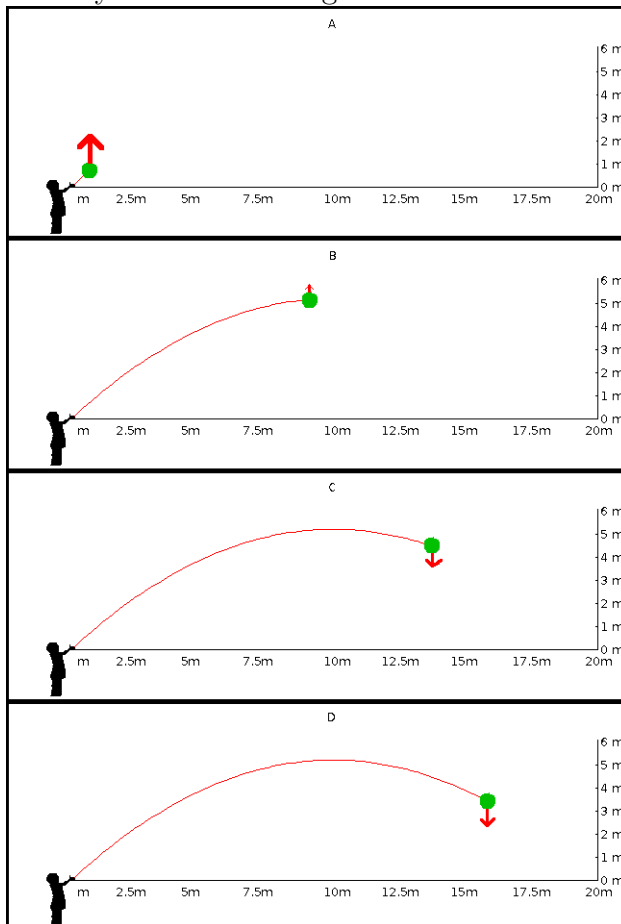
7. A boy has tossed an apple up and to the right. Consider the trajectory of the apple shown in the figure when answering the following questions. Ignore the effects of air resistance.



- a. Where is the magnitude of the vertical velocity largest? (*illustrations*)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal
- b. Where is the magnitude of the horizontal velocity largest? (*illustrations*)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal
- c. Where is the magnitude of vertical velocity smallest? (*illustrations*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

8. A boy has tossed an apple up and to the right (just like on the previous page). Consider these figures below when answering the following questions. The arrows represent the velocity at that time. Ignore the effects of air resistance.



- a. Between time A and time B, what is the direction of vertical **velocity**? (*illustrations*)
- up
 - down
 - left
 - right

- the apple isn't moving vertically
- b. Between time A and time B, what is the direction of vertical **acceleration**? (*illustrations*)
- up
 - down
 - left
 - right
 - the apple isn't moving vertically
- c. Between time C and time D, what is the direction of vertical **velocity**? (*illustrations*)
- up
 - down
 - left
 - right
 - the apple isn't moving vertically
- d. Between time C and time D, what is the direction of vertical **acceleration**? (*illustrations*)
- up
 - down
 - left
 - right
 - the apple isn't moving vertically
- e. Between time A and time B, what is the direction of horizontal **velocity**? (*illustrations*)
- up
 - down
 - left
 - right
 - the apple isn't moving horizontally
- f. Between time A and time B, what is the direction of horizontal **acceleration**?

(illustrations)

- up
- down
- left
- right
- the apple isn't moving horizontally

9. Which of the following is the definition of acceleration? *(conceptual)*

- change in velocity over change in time
- change in speed over change in time
- change in displacement over change in time
- None of the above.

10. Is acceleration a vector or scalar variable? *(conceptual)*

- vector
- scalar

D.2 TEST B

This section shows the test questions for test B.

1. A toy car runs off the edge of a table that is 1.225 m high. The car lands 0.40 m from the base of the table.

a. Which of the equations below would you use to determine how long it took for the car to fall? *(numeric)*

- $d_y = v_{0y} * t + 1/2 * g * t^2$ (in Andes: $d_y = v_{0_y} * t + 1/2 * g * t^2$)
- $v_y^2 = v_{0y}^2 + 2 * g * d_y$ (in Andes: $v_y^2 = v_{0_y}^2 + 2 * g * d_y$)
- $v_y = v_{0y} + g * t$ (in Andes: $v_y = v_{0_y} + g * t$)
- $g = (v_y - v_{0y})/t$ (in Andes: $g = (v_y - v_{0_y}) / t$)

b. Which number would you use for v_{0y} ? *(numeric)*

- 0 m/s
- 1.225 m

- 0.40 m
- 9.8 m/s^2
- That value is not known for this problem.

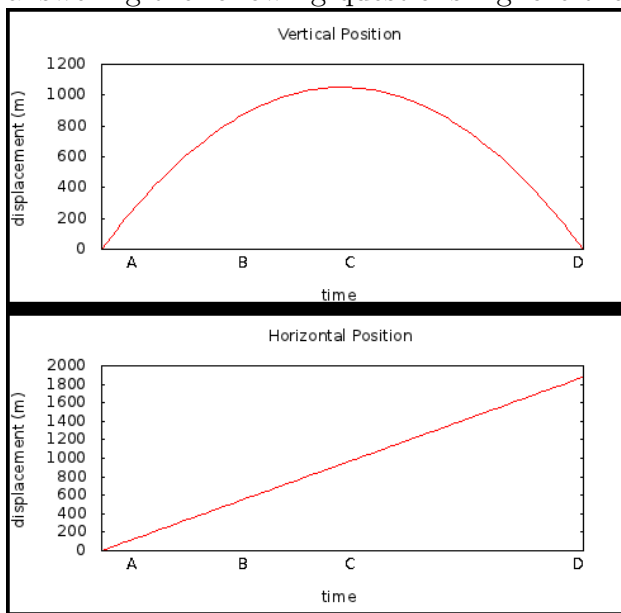
c. Assuming you have solved for time already (in part a), which of the equations below would you use to determine how fast the orange was rolling as it rolled off the edge? (*numeric*)

- $v_y = v_{0y} + g * t$ (in Andes: $v_y = v0_y + g * t$)
- $v_x^2 = v_{0x}^2 + 2 * a_x * d_x$ (in Andes: $v_x^2 = v0_x^2 + 2 * a_x * d_x$)
- $d_y = v_{0y} * t + 1/2 * g * t^2$ (in Andes: $d_y = v0_y * t + 1/2 * g * t^2$)
- $d_x = v_{0x} * t$ (in Andes: $d_x = v0_x * t$)

d. What physics quantity represents change in velocity over time? (*conceptual*)

- displacement
- acceleration
- velocity
- None of the above

2. A cannonball was shot over a wall towards a castle. The two graphs below show the vertical and horizontal positions of the ball over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.



a. Where is the magnitude of the vertical velocity largest? (*graphs*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

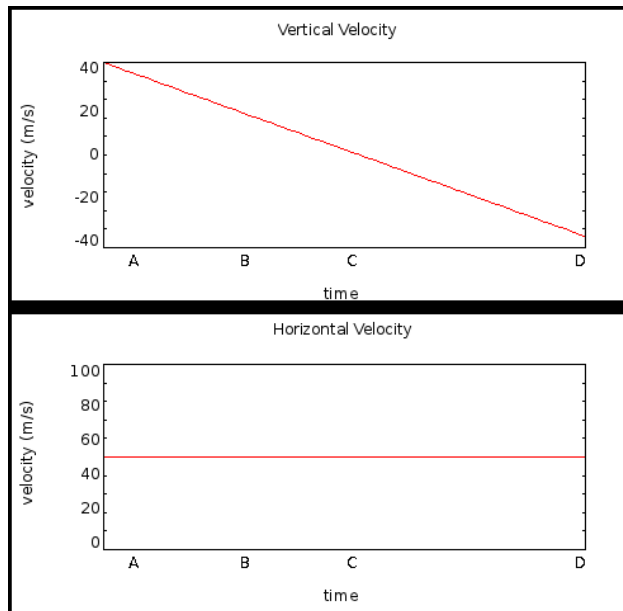
b. Where is the magnitude of the horizontal velocity largest? (*graphs*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

c. Where is the magnitude of vertical velocity smallest? (*graphs*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

3. A cannonball was shot over a wall towards a castle to the right (just like on the previous page). Now, the graphs below show the vertical and horizontal **velocities** over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.



- a. Between time A and time B, what is the direction of vertical **velocity**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving vertically
- b. Between time A and time B, what is the direction of vertical **acceleration**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving vertically
- c. Between time C and time D, what is the direction of vertical **velocity**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving vertically
- d. Between time C and time D, what is the direction of vertical **acceleration**? (*graphs*)

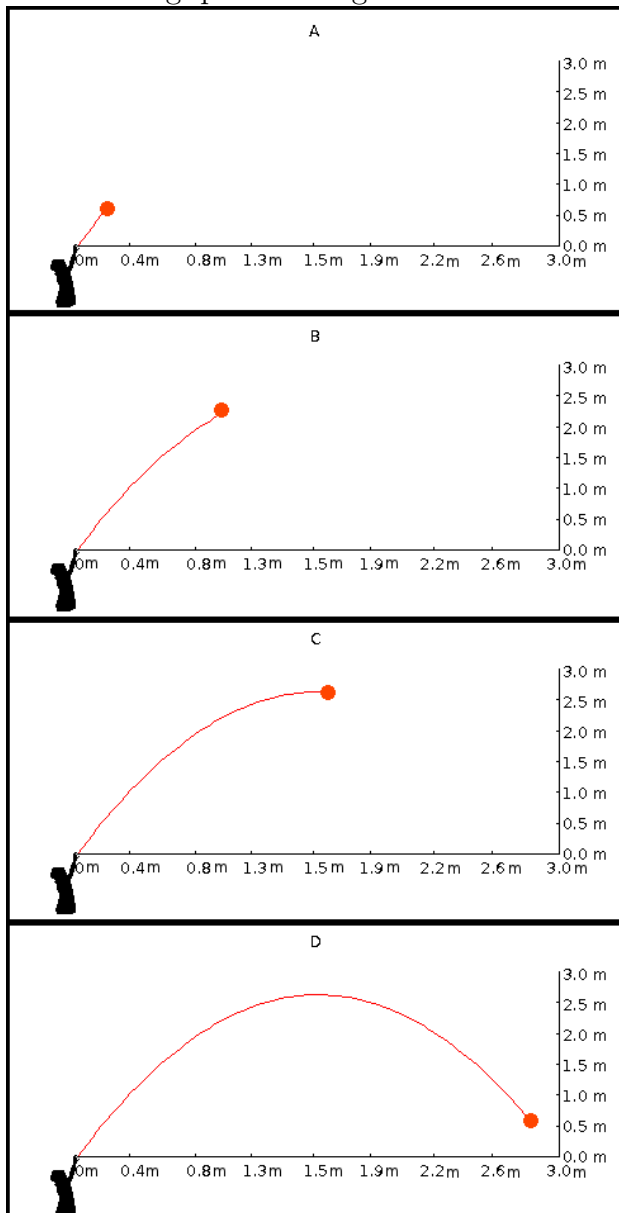
- up
 - down
 - left
 - right
 - the ball isn't moving vertically
- e. Between time A and time B, what is the direction of horizontal **velocity**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving horizontally
- f. Between time A and time B, what is the direction of horizontal **acceleration**? (*graphs*)
- up
 - down
 - left
 - right
 - the ball isn't moving horizontally
4. A child tosses a bouncy ball straight into the air.
- a. After it leaves his hand, **on the way up**, what is the direction of velocity? (*conceptual*)
- Up
 - Down
 - Left
 - Right
 - There is not enough information to know.
- b. After it leaves his hand, **on the way up**, what is the direction of acceleration? (*conceptual*)
- Up
 - Down

- Left
 - Right
 - There is not enough information to know.
- c. **On its way back down**, what is the direction of velocity? (*conceptual*)
- Up
 - Down
 - Left
 - Right
 - There is not enough information to know.
- d. **On its way back down**, what is the direction of acceleration? (*conceptual*)
- Up
 - Down
 - Left
 - Right
 - There is not enough information to know.
5. Suppose that a hot air balloonist drops an apple from rest over the side while the balloon is hovering stationary. By how much would the apple's speed increase with each second of fall? Ignore air resistance. (*numeric*)
- 19.6 m/s
 - 9.8 m/s
 - 4.9 m/s
 - 0.0 m/s, the speed is constant
 - None of the above.
6. A train is traveling north towards a train station.
- a. As it arrives at the station, it slows to a stop. What is the direction of acceleration? (*conceptual*)
- North
 - South
 - There is no acceleration
 - None of the above

b. When the train departs the station, it heads north to the next station. What is the direction of acceleration as it speeds up? (*conceptual*)

- North
- South
- There is no acceleration
- None of the above

7. A child tosses a ball up and to the right. The two graphs shown below show the vertical and horizontal positions of the ball over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.



a. Where is the magnitude of the vertical velocity largest? (*illustrations*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

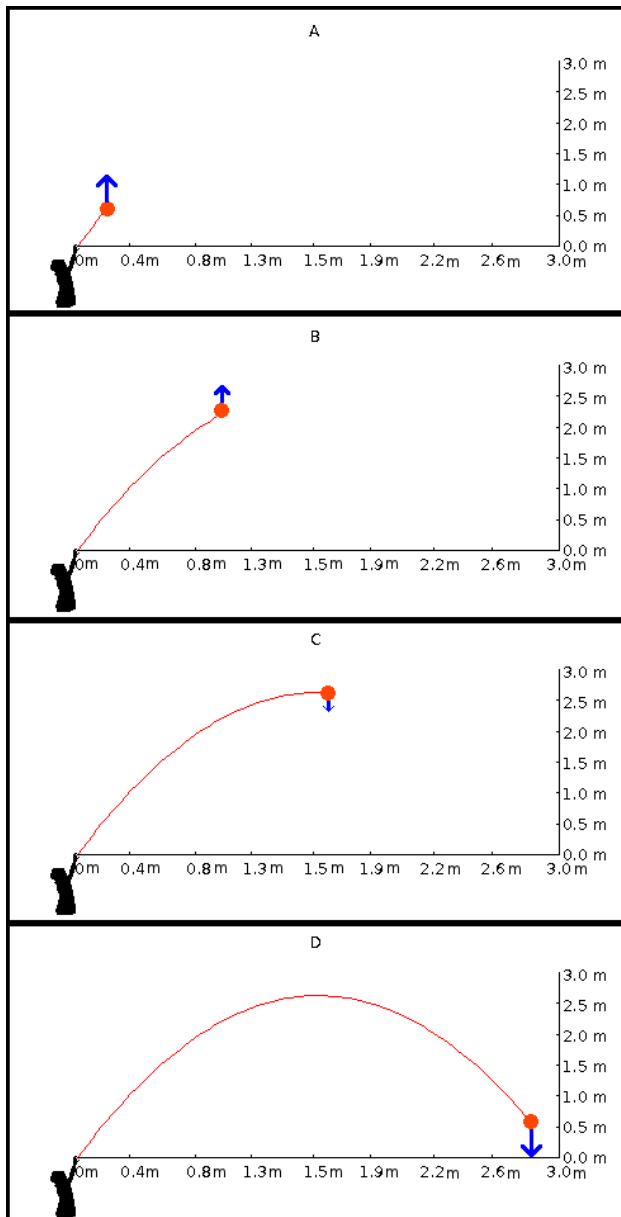
b. Where is the magnitude of the horizontal velocity largest? (*illustrations*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

c. Where is the magnitude of vertical velocity smallest? (*illustrations*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

8. A child tosses a ball up and to the right (just like on the previous page). Consider these figures below when answering the following questions. The arrows represent the velocity at that time. Ignore the effects of air resistance.



a. Between time A and time B, what is the direction of vertical **velocity**? (*illustrations*)

- up
- down
- left
- right
- the apple isn't moving vertically

b. Between time A and time B, what is the direction of vertical **acceleration**? (*illus-*

- trations)*
- up
 - down
 - left
 - right
 - the apple isn't moving vertically
- c. Between time C and time D, what is the direction of vertical **velocity**? (*illustrations)*
- up
 - down
 - left
 - right
 - the apple isn't moving vertically
- d. Between time C and time D, what is the direction of vertical **acceleration**? (*illustrations)*
- up
 - down
 - left
 - right
 - the apple isn't moving vertically
- e. Between time A and time B, what is the direction of horizontal **velocity**? (*illustrations)*
- up
 - down
 - left
 - right
 - the apple isn't moving horizontally
- f. Between time A and time B, what is the direction of horizontal **acceleration**? (*illustrations)*
- up

- down
 - left
 - right
 - the apple isn't moving horizontally
9. Which of the following is the definition of acceleration? (*conceptual*)
- change in velocity over change in time
 - change in speed over change in time
 - change in displacement over change in time
 - None of the above.
10. Is acceleration a vector or scalar variable? (*conceptual*)
- vector
 - scalar

APPENDIX E

FIRST STUDY PROBLEMS AND REFLECTION QUESTIONS

E.1 PROBLEM 1: KANGAROO

Problem Statement: A kangaroo can jump about 1.50 m straight up. What is the magnitude of the take-off velocity?

Reflection Questions:

1. At what time was the kangaroo's velocity greatest?
2. So we know when the kangaroo is moving fastest, but why is it slowing down?
3. What was the direction of acceleration?

The image shown during the problem-solving for the illustration condition can be found in [E1a](#); the graph version can be found in [E1b](#).

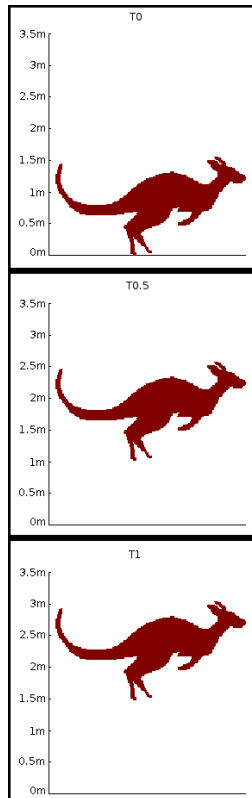
E.2 PROBLEM 2: ARROW

Problem Statement: Suppose you aim a bow horizontally, directly at the center of a target 25.0 m away from you. If the speed of the arrow is 60 m/s, how far from the center of the target will it strike the target? That is, find the vertical displacement of the arrow. Assume there is no air friction.

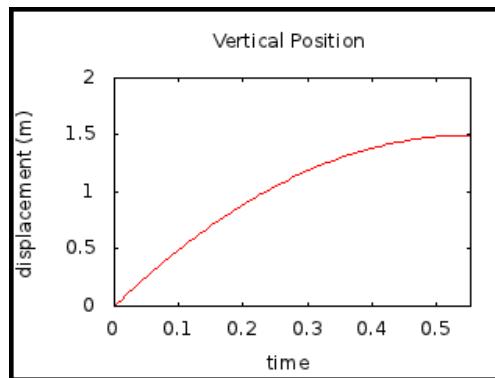
Reflection Questions:

1. Why did the arrow hit below the target?
2. How can the archer hit the target?
3. Does the arrow experience any acceleration in the x-direction while in flight?

The image shown during the problem-solving for the illustration condition can be found in [E2a](#); the graph version can be found in [E2b](#).



(a) An illustration of the situation, shown only to students in the illustration condition.



(b) A graph of the situation, shown only to students in the graph condition.

Figure E1: The two possible visual representations for the kangaroo problem are shown.

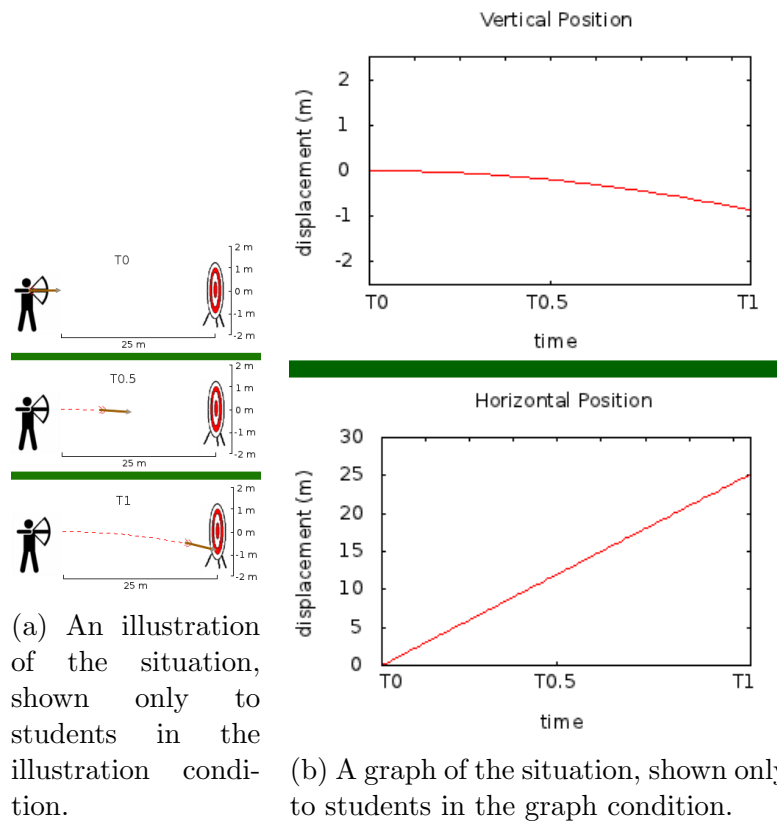


Figure E2: The two possible visual representations for the kangaroo problem are shown.

APPENDIX F

SUPPLEMENTAL RESULT ANALYSIS FOR STUDY 1

F.1 TABLES FOR PRETEST SCORE AND STUDY CONDITION ANOVAS

Table F1: Raw learning gain pretest-condition ANOVA results for the five test score measures. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	Pretest=High		Pretest=Low		Interaction p-val
	Illus.	Graph	Illus.	Graph	
Overall	0.129	0.161	0.382	0.151	0.123
Numeric	0.100	0.050	0.400	0.067	0.438
Conceptual	0.135	0.183	0.379	0.167	0.090
Illustrations	0.167	0.222	365	0.222	0.228
Graphs	0.222	0.194	0.413	0.185	0.105

Table F2: Normalized learning gain pretest-condition ANOVA results for the five test score measures. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	Pretest=High		Pretest=Low		Interaction p-val
	Illus.	Graph	Illus.	Graph	
Overall	0.556	0.668	0.614	0.316	0.094
Numeric	0.500	0.500	0.462	0.093	0.427
Conceptual	0.571	0.701	0.648	0.365	0.062
Illustrations	0.667	0.854	0.607	0.446	0.049
Graphs	0.750	0.688	0.612	0.329	0.013

Table F3: Raw learning gain pretest-condition ANOVA results for the five test score measures. All 29 subjects are included in this table. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	Pretest=High		Pretest=Low		Interaction p-val
	Illus.	Graph	Illus.	Graph	
Overall	-0.032	0.110	0.335	0.135	0.121
Problem-Solving	-0.067	0.000	0.350	0.060	0.076
Conceptual	-0.026	0.131	0.332	0.150	0.149
Illustrations	0.019	0.156	0.306	0.200	0.173
Graphs	0.019	0.156	0.375	0.189	0.183

Table F4: Normalized learning gain pretest-condition ANOVA results for the five test score measures. All 29 subjects are included in this table. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	Pretest=High		Pretest=Low		Interaction p-val
	Illus.	Graph	Illus.	Graph	
Overall	0.052	0.334	0.537	0.284	0.296
Problem-Solving	0.000	0.200	0.404	0.083	0.281
Conceptual	0.065	0.361	0.567	0.329	0.311
Illustrations	0.222	0.683	0.506	0.402	0.098
Graphs	0.194	0.550	0.560	0.330	0.173

F.2 SPATIAL REASONING AND STUDY CONDITION ANOVAS TABLES

Table F5: Raw learning gain spatial reasoning-condition ANOVA results for the five test score measures. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	Spatial=High		Spatial=Low		Interaction p-val
	Illus.	Graph	Illus.	Graph	
Overall	0.371	0.188	0.237	0.124	0.190
Problem-Solving	0.333	0.067	0.333	0.057	0.531
Conceptual	0.378	0.212	0.218	0.137	0.147
Illustrations	0.370	0.259	0.222	0.190	0.440
Graphs	0.444	0.241	0.222	0.143	0.510

Table F6: Normalized learning gain spatial reasoning-condition ANOVA results for the five test score measures. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	Spatial=High		Spatial=Low		Interaction p-val
	Illus.	Graph	Illus.	Graph	
Overall	0.697	0.598	0.410	0.275	0.261
Problem-Solving	0.522	0.389	0.367	0.071	0.609
Conceptual	0.736	0.646	0.421	0.316	0.220
Illustrations	0.736	0.769	0.389	0.402	0.378
Graphs	0.775	0.673	0.378	0.240	0.184

Table F7: Raw learning gain spatial reasoning-condition ANOVA results for the five test score measures. All 29 subjects are included in this table. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	Spatial=High		Spatial=Low		Interaction p-val
	Illus.	Graph	Illus.	Graph	
Overall	0.211	0.147	0.116	0.109	0.481
Problem-Solving	0.178	0.029	0.160	0.050	0.389
Conceptual	0.218	0.170	0.108	0.120	0.521
Illustrations	0.210	0.206	0.133	0.167	0.601
Graphs	0.284	0.206	0.111	0.153	0.261

Table F8: Normalized learning gain spatial reasoning-condition ANOVA results for the five test score measures. All 29 subjects are included in this table. The scores are percentages out of the total number of questions for that measure (e.g. out of nine for graphs).

Measure	Spatial=High		Spatial=Low		Interaction p-val
	Illus.	Graph	Illus.	Graph	
Overall	0.465	0.370	0.086	0.241	0.455
Problem-Solving	0.348	0.190	0.020	0.062	0.416
Conceptual	0.490	0.411	0.103	0.277	0.483
Illustrations	0.469	0.660	0.233	0.352	0.599
Graphs	0.539	0.577	0.160	0.251	0.675

APPENDIX G

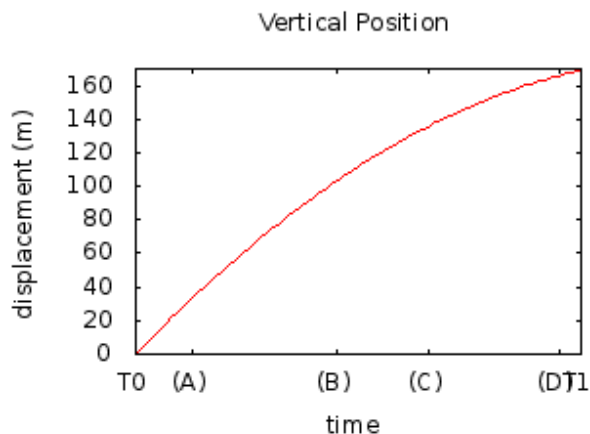
TEST QUESTIONS FOR THE SECOND STUDY

This appendix shows the two tests for the second study, described in Chapter 6. Each question is also tagged for applicable knowledge components and is indicated in italics after the question.

G.1 TEST A

This section shows the test questions for test A.

1. Two sailors are stranded at sea. To signal an approaching plane, they fire a flare straight into the air. The graph shown below has the vertical position of the flare. Consider this graph when answering the following questions. Ignore the effects of air resistance.



a. Where is the magnitude of the vertical position largest?

- Position A
- Position B
- Position C
- Position D
- All positions equal

b. Where is the magnitude of vertical position smallest?

- Position A
- Position B
- Position C
- Position D
- All positions equal

c. Where is the magnitude of the vertical velocity largest? (*KC28*)

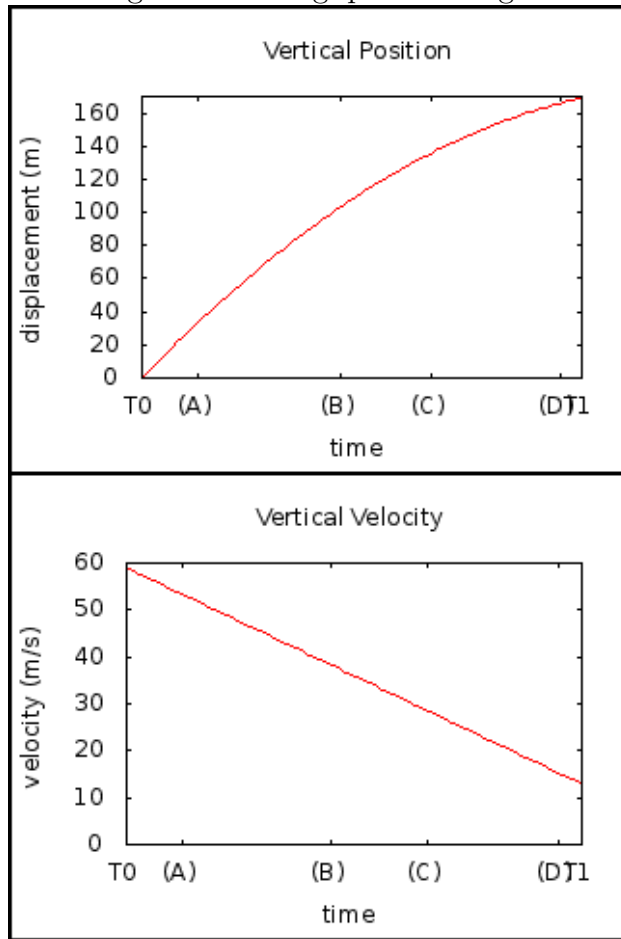
- Position A
- Position B
- Position C
- Position D
- All positions equal

d. Where is the magnitude of vertical velocity smallest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions equal

2. A flare is shot straight up into the air (as on the previous page). Now, the graphs below show the vertical position and vertical velocity over time. Consider these graphs when

answering the following questions. Ignore the effects of air resistance.



- a. Between time A and time B, what is the direction of vertical velocity?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- b. Between time A and time B, what is the direction of vertical acceleration? (*KC15*, *KC23*)
- up
 - down
 - left
 - right

- it isn't moving vertically

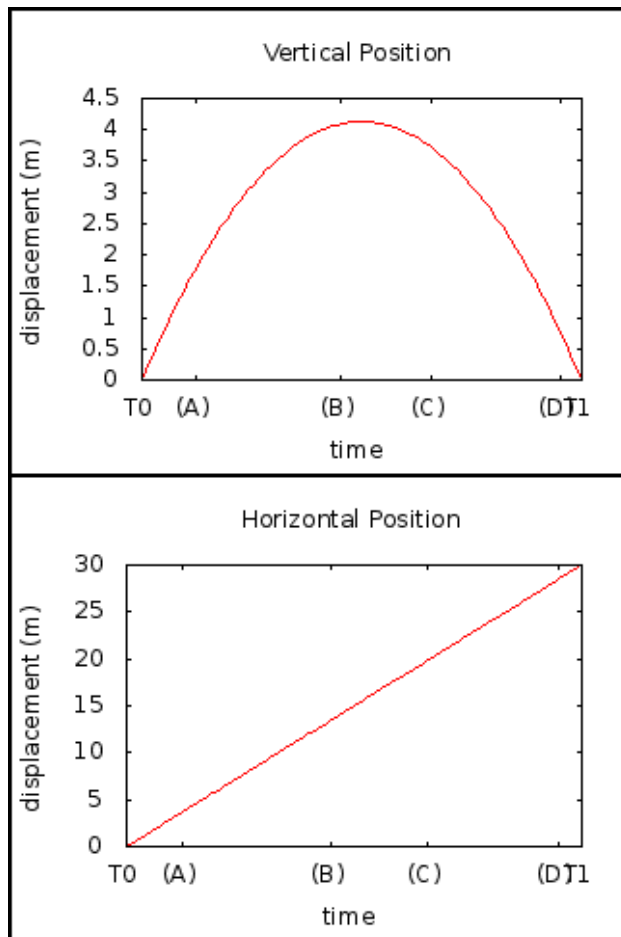
c. Between time C and time D, what is the direction of vertical velocity?

- up
- down
- left
- right
- it isn't moving vertically

d. Between time A and time B, what is the direction of vertical acceleration? (*KC15*, *KC23*)

- up
- down
- left
- right
- it isn't moving vertically

3. A soccer player kicks a ball into the air so that it goes over opposing player and lands near a teammate down the field. The two graphs shown below show the vertical and horizontal positions of the ball over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.

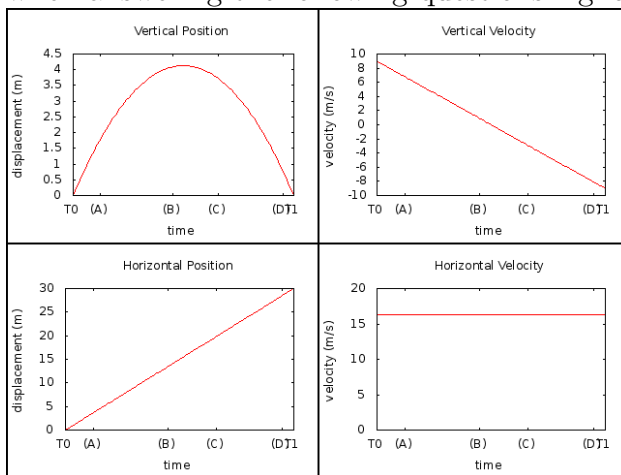


- a. Where is the magnitude of the vertical velocity largest? (KC28)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal
- b. Where is the magnitude of the horizontal velocity largest? (KC28)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal

c. Where is the magnitude of vertical velocity smallest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

4. A soccer player kicks a ball into the air so that it goes over opposing player and lands near a teammate down the field (just like on the previous page). Now, the graphs below show the vertical and horizontal positions and velocities over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.



a. Between time A and time B, what is the direction of vertical **velocity**?

- up
- down
- left
- right
- it isn't moving vertically

b. Between time A and time B, what is the direction of vertical **acceleration**? (*KC15, KC23*)

- up

- down
 - left
 - right
 - it isn't moving vertically
- c. Between time C and time D, what is the direction of vertical **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- d. Between time C and time D, what is the direction of vertical **acceleration**? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically
- e. Between time A and time B, what is the direction of horizontal **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving horizontally
- f. Between time A and time B, what is the direction of horizontal **acceleration**? (*KC15, KC23*)
- up

- down
- left
- right
- it isn't moving horizontally

5. Suppose someone dropped a ball from the top of a tall building (e.g. the Cathedral of Learning). By how much would the balls speed increase with each second of fall? Ignore air resistance. (*KC24*)

- 19.6 m/s
- 9.8 m/s
- 4.9 m/s
- 0.0 m/s, the speed is constant
- None of the above.

6. During baseball practice, an outfielder throws the baseball perfectly horizontally to see how far it would go before it lands.

a. When did the baseball begin to fall? (*KC24*)

- The moment it left his hand
- A few moments after it left his hand
- It depends on how fast it was thrown
- It never falls

b. If the ball hit the ground in 2.7 seconds, would it land in more or less time if he had thrown it softer? (*KC32*)

- More time
- Less time
- Same amount of time
- It is not possible to tell with the information given

c. If the pitcher were standing on the top of the Grand Canyon and threw the baseball,

when would the ball be falling straight down? (*KC25*)

- Near the end of the fall/throw
- Itll never fall straight down
- It depends on how fast it was thrown

d. In what direction was gravity working? (*KC24*)

- Horizontally
- Vertically
- Both horizontally and vertically
- There was no gravity in this situation

7. A woman tosses her keys straight up into the air.

a. **On the way up**, after it leaves her hand, what is the direction of velocity? (*KC28*)

- up
- down
- left
- right
- There is not enough information to know.

b. **On the way up**, after it leaves her hand, what is the direction of acceleration?
(*KC15*)

- up
- down
- left
- right
- There is not enough information to know.

c. **On the way back down**, what is the direction of velocity?(*KC28*)

- up
- down

- left
- right
- There is not enough information to know.

d. **On the way back down**, what is the direction of acceleration? (*KC15*)

- up
- down
- left
- right
- There is not enough information to know.

e. **On the way up**, when were the keys moving fastest? (*KC28*)

- Just after leaving her hand
- Around the midpoint of the ascent
- At the top of the toss
- There is not enough information to know

8. An orange rolls off the edge of a table and lands on the floor. The table is 1.50 m high and the orange lands 0.6 m from the base of the table.

a. Which of the equations below would you use to determine how long it took for the orange to fall? (*KC32*)

· $d_y = v_{0y} * t + 1/2 * g * t^2$ (in Andes: $d_y = v_{0_y} * t + 1/2 * g * t^2$)

· $v_y^2 = v_{0y}^2 + 2 * g * d_y$ (in Andes: $v_y^2 = v_{0_y}^2 + 2 * g * d_y$)

· $v_y = v_{0y} + g * t$ (in Andes: $v_y = v_{0_y} + g * t$)

· $g = (v_y - v_{0y})/t$ (in Andes: $g = (v_y - v_{0_y}) / t$)

b. Which number would you use for v_{0y} ?

- 0 m/s
- 1.50 m
- 0.6 m

- 9.8 m/s^2
- That value is not known for this problem.

c. Assuming you have solved for time already (in part a), which of the equations below would you use to determine how fast the orange was rolling as it rolled off the edge?

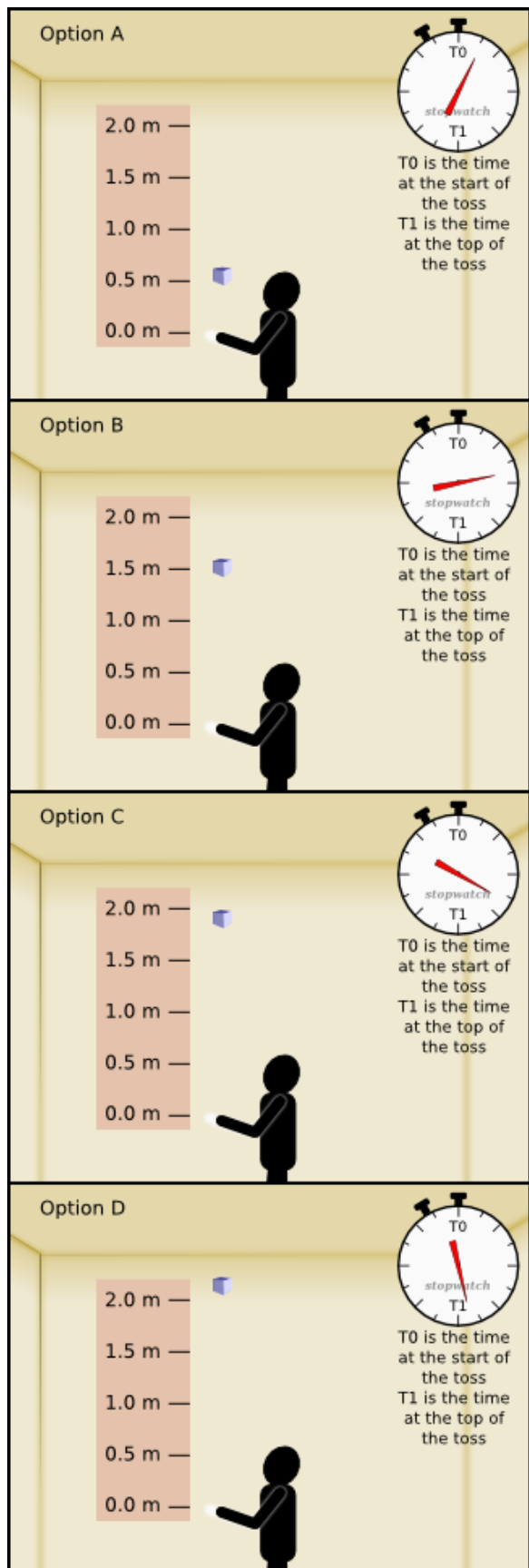
(*KC32, KC25*)

- $v_y = v_{0y} + g * t$ (in Andes: $v_y = v_{0y} + g * t$)
- $v_x^2 = v_{0x}^2 + 2 * a_x * d_x$ (in Andes: $v_x^2 = v_{0x}^2 + 2 * a_x * d_x$)
- $d_y = v_{0y} * t + 1/2 * g * t^2$ (in Andes: $d_y = v_{0y} * t + 1/2 * g * t^2$)
- $d_x = v_{0x} * t$ (in Andes: $d_x = v_{0x} * t$)

d. When does the orange begin falling? (*KC24*)

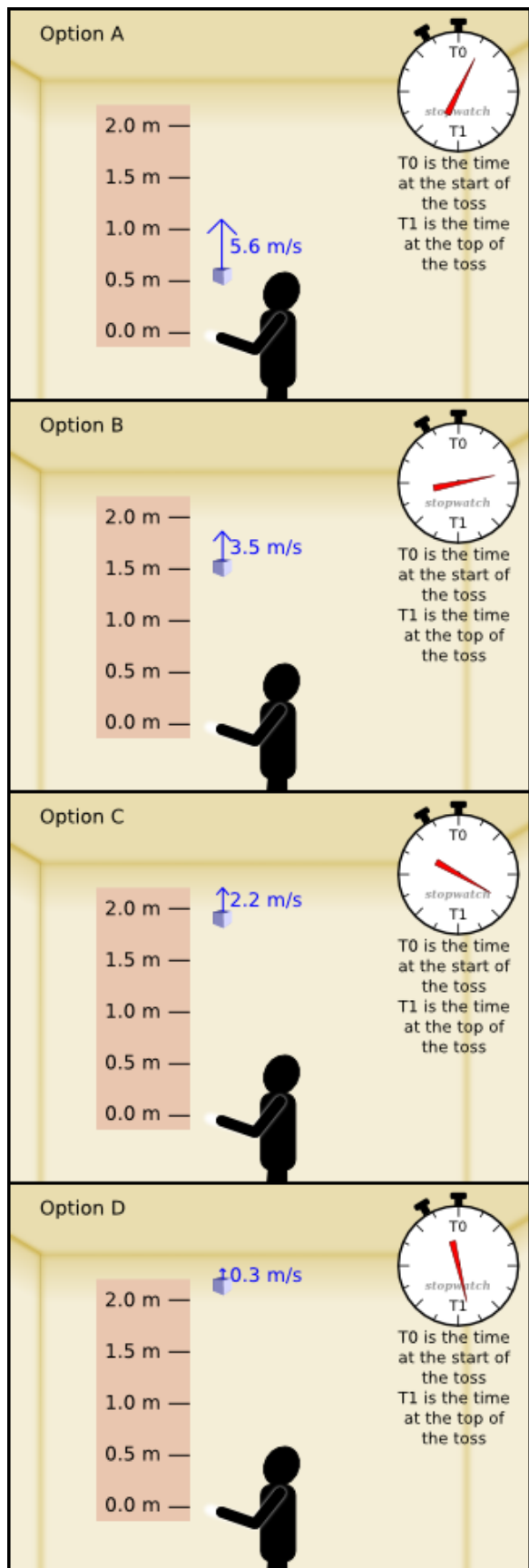
- Immediately after it leaves the table
- A few moments after it leaves the table
- It depends on how fast it was rolling
- It never falls

9. A young child was throwing a block straight up into the air to see if she could hit the ceiling. Consider the trajectory of the block shown in the figure when answering the following questions. Ignore the effects of air resistance.



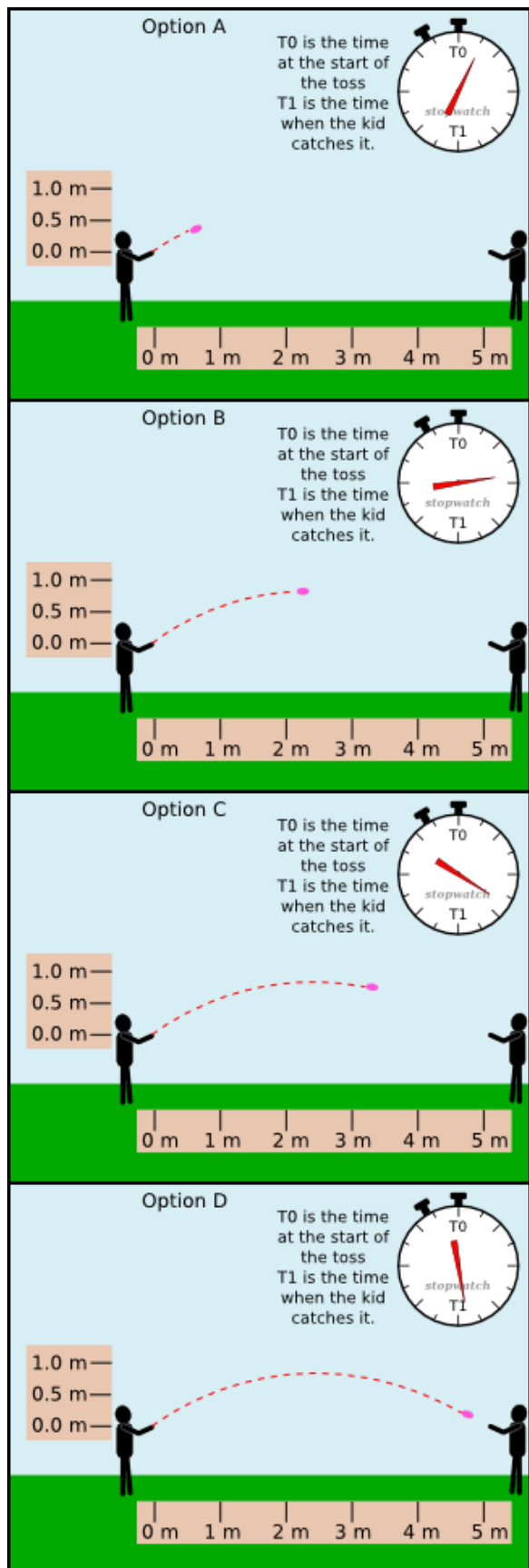
- a. Where is the magnitude of the vertical position largest?
- Position A
 - Position B
 - Position C
 - Position D
 - All positions equal
- b. Where is the magnitude of vertical position smallest?
- Position A
 - Position B
 - Position C
 - Position D
 - All positions equal
- c. Where is the magnitude of the vertical velocity largest? (*KC28*)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions equal
- d. Where is the magnitude of vertical velocity smallest? (*KC28*)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions equal

10. A child is tossing a block up into the air (just like on the previous page). Consider these figures below when answering the following questions. Ignore the effects of air resistance.



- a. Between time A and time B, what is the direction of vertical velocity?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- b. Between time A and time B, what is the direction of vertical acceleration? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically
- c. Between time C and time D, what is the direction of vertical velocity?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- d. Between time A and time B, what is the direction of vertical acceleration? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically

11. Two kids are playing a game of water balloon catch. Consider the trajectory of the water balloon shown in the figures when answering the following questions. Ignore the effects of air resistance.



a. Where is the magnitude of the vertical velocity largest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

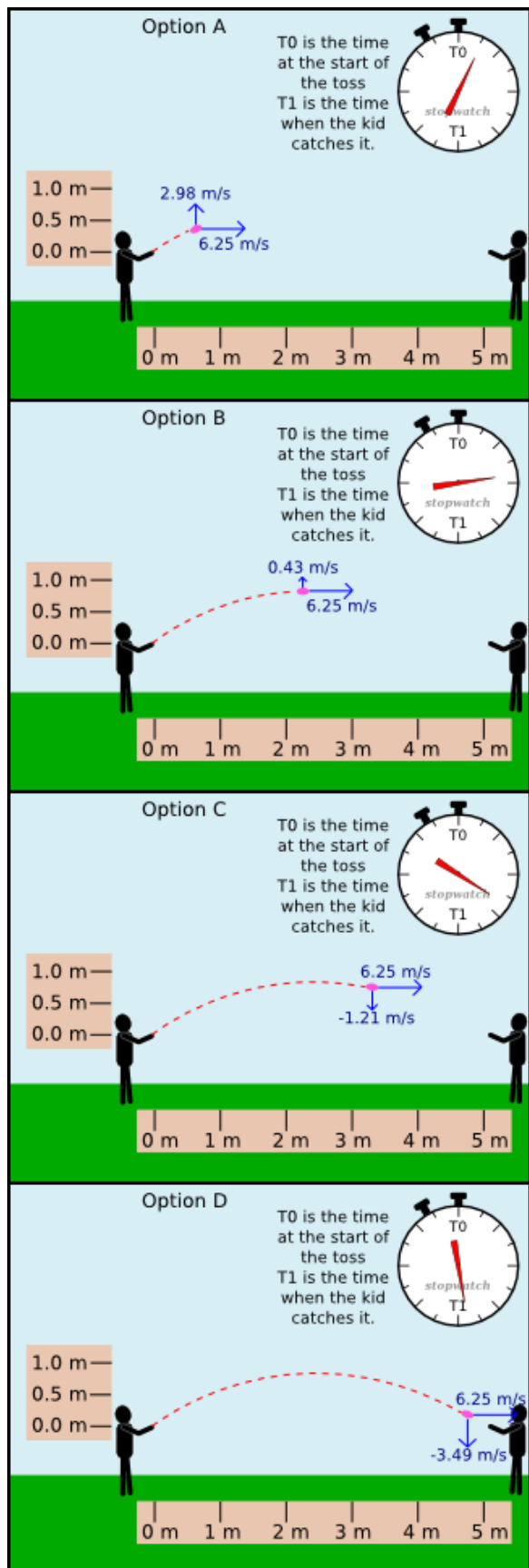
b. Where is the magnitude of the horizontal velocity largest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

c. Where is the magnitude of vertical velocity smallest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

12. Two kids are playing a game of water balloon catch (just like on the previous page). Consider these figures below when answering the following questions. Ignore the effects of air resistance.



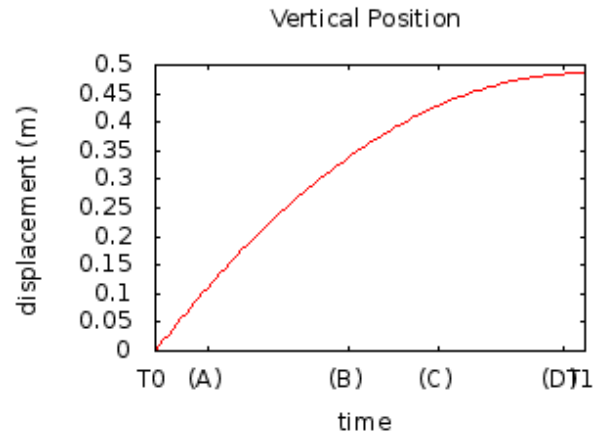
- a. Between time A and time B, what is the direction of vertical **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- b. Between time A and time B, what is the direction of vertical **acceleration**? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically
- c. Between time C and time D, what is the direction of vertical **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- d. Between time C and time D, what is the direction of vertical **acceleration**? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically

- e. Between time A and time B, what is the direction of horizontal **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving horizontally
- f. Between time A and time B, what is the direction of horizontal **acceleration**?
(*KC15*, *KC23*)
- up
 - down
 - left
 - right
 - it isn't moving horizontally

G.2 TEST B

This section shows the test questions for test B.

1. To entertain his baby, a father playfully tosses the child straight up. The graph shown below has the vertical position of the baby. Consider this graph when answering the following questions. Ignore the effects of air resistance.



- a. Where is the magnitude of the vertical position largest?
 - Position A
 - Position B
 - Position C
 - Position D
 - All positions equal

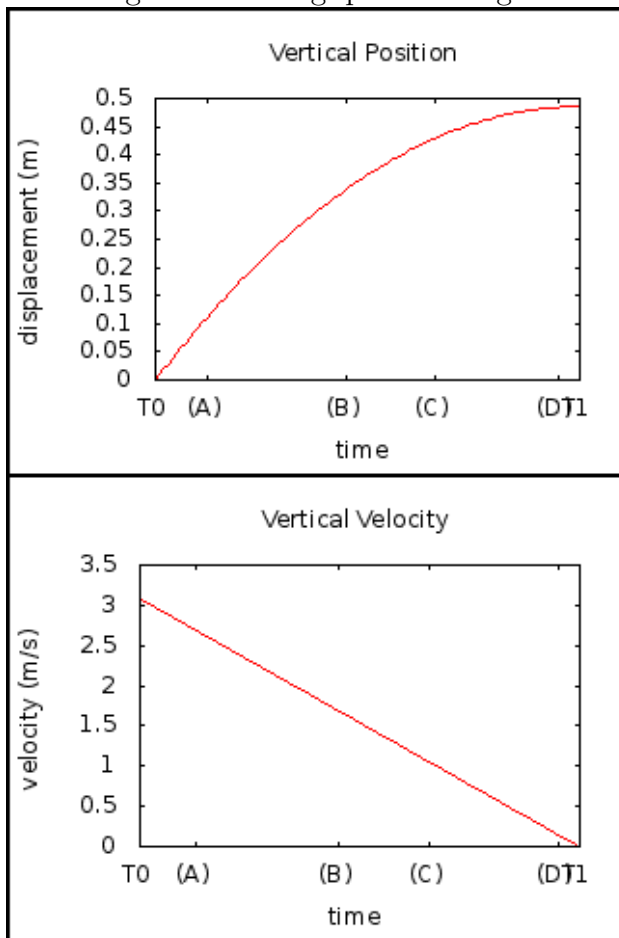
- b. Where is the magnitude of vertical position smallest?
 - Position A
 - Position B
 - Position C
 - Position D
 - All positions equal

- c. Where is the magnitude of the vertical velocity largest? (*KC28*)
 - Position A
 - Position B
 - Position C
 - Position D
 - All positions equal

d. Where is the magnitude of vertical velocity smallest? (KC28)

- Position A
- Position B
- Position C
- Position D
- All positions equal

2. A baby is tossed up into the air (as on the previous page). Now, the graphs below show the vertical position and vertical velocity over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.



a. Between time A and time B, what is the direction of vertical velocity?

- up
- down
- left

- right
- it isn't moving vertically

b. Between time A and time B, what is the direction of vertical acceleration? (*KC15, KC23*)

- up
- down
- left
- right
- it isn't moving vertically

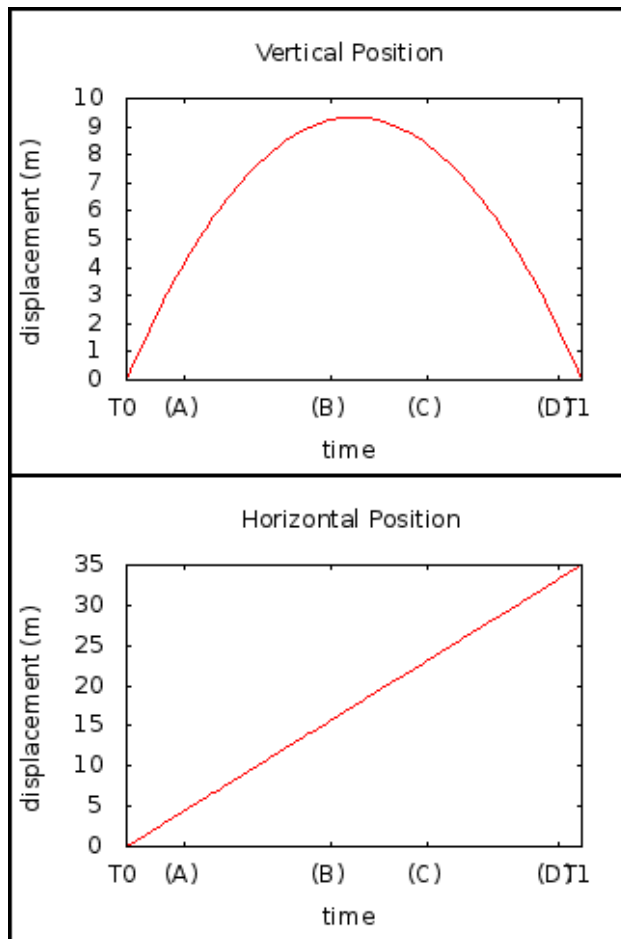
c. Between time C and time D, what is the direction of vertical velocity?

- up
- down
- left
- right
- it isn't moving vertically

d. Between time A and time B, what is the direction of vertical acceleration? (*KC15, KC23*)

- up
- down
- left
- right
- it isn't moving vertically

3. A medieval army catapults an explosive device at an approaching army. The two graphs shown below show the vertical and horizontal positions of the explosive device over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.

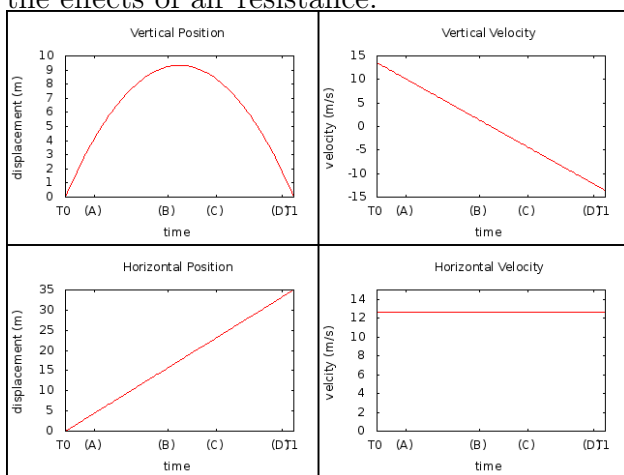


- a. Where is the magnitude of the vertical velocity largest? (KC28)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal
- b. Where is the magnitude of the horizontal velocity largest? (KC28)
- Position A
 - Position B
 - Position C
 - Position D
 - All positions are equal

c. Where is the magnitude of vertical velocity smallest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

4. A medieval army catapults an explosive device at an approaching army (just like the previous page). Now, the graphs below show the vertical and horizontal positions and velocities over time. Consider these graphs when answering the following questions. Ignore the effects of air resistance.



a. Between time A and time B, what is the direction of vertical **velocity**?

- up
- down
- left
- right
- it isn't moving vertically

b. Between time A and time B, what is the direction of vertical **acceleration**? (*KC15, KC23*)

- up

- down
 - left
 - right
 - it isn't moving vertically
- c. Between time C and time D, what is the direction of vertical **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- d. Between time C and time D, what is the direction of vertical **acceleration**? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically
- e. Between time A and time B, what is the direction of horizontal **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving horizontally
- f. Between time A and time B, what is the direction of horizontal **acceleration**? (*KC15, KC23*)
- up

- down
- left
- right
- it isn't moving horizontally

5. Suppose that a hot air balloonist drops an apple from rest over the side while the balloon is hovering stationary. By how much would the apple's speed increase with each second of fall? Ignore air resistance. (*KC24*)

- 19.6 m/s
- 9.8 m/s
- 4.9 m/s
- 0.0 m/s, the speed is constant
- None of the above.

6. During football practice, the quarterback throws a football down the field to see how far it would go before it lands. He threw the football exactly horizontally.

a. When did the football begin to fall? (*KC24*)

- The moment it left his hand
- A few moments after it left his hand
- It depends on how fast it was thrown
- It never falls

b. If the ball hit the ground in 2.7 seconds, would it land in more or less time if he had thrown it softer? (*KC32*)

- More time
- Less time
- Same amount of time
- It is not possible to tell with the information given

c. If the quarterback were standing at the very top of the football stadium and threw

the football, when would the ball be falling straight down? (*KC25*)

- Near the end of the fall/throw
- Itll never fall straight down
- It depends on how fast it was thrown

d. In what direction was gravity working? (*KC24*)

- Horizontally
- Vertically
- Both horizontally and vertically
- There was no gravity in this situation

7. A child tosses a bouncy ball straight into the air.

a. After it leaves his hand,**on the way up**, what is the direction of its velocity?

(*KC28*)

- up
- down
- left
- right
- There is not enough information to know.

b. After it leaves his hand,**on the way up**, what is the direction of its acceleration?

(*KC15*)

- up
- down
- left
- right
- There is not enough information to know.

c. **On the way back down**, what is the direction of velocity? (*KC28*)

- up

- down
- left
- right
- There is not enough information to know.

d. **On the way back down**, what is the direction of acceleration? (*KC15*)

- up
- down
- left
- right
- There is not enough information to know.

e. **On the way up**, when were the keys moving fastest? (*KC28*)

- Just after leaving his hand
- Around the midpoint of the ascent
- At the top of the toss
- There is not enough information to know

8. A toy car runs off the edge of a table that is 1.225 m high. The car lands 0.40 m from the base of the table.

a. Which of the equations below would you use to determine how long it took for the car to fall? (*KC32*)

· $d_y = v_{0y} * t + 1/2 * g * t^2$ (in Andes: $d_y = v_{0_y} * t + 1/2 * g * t^2$)

· $v_y^2 = v_{0y}^2 + 2 * g * d_y$ (in Andes: $v_y^2 = v_{0_y}^2 + 2 * g * d_y$)

· $v_y = v_{0y} + g * t$ (in Andes: $v_y = v_{0_y} + g * t$)

· $g = (v_y - v_{0y})/t$ (in Andes: $g = (v_y - v_{0_y}) / t$)

b. Which number would you use for v_{0y} ?

- 0 m/s
- 1.50 m

- 0.6 m
- 9.8 m/s^2
- That value is not known for this problem.

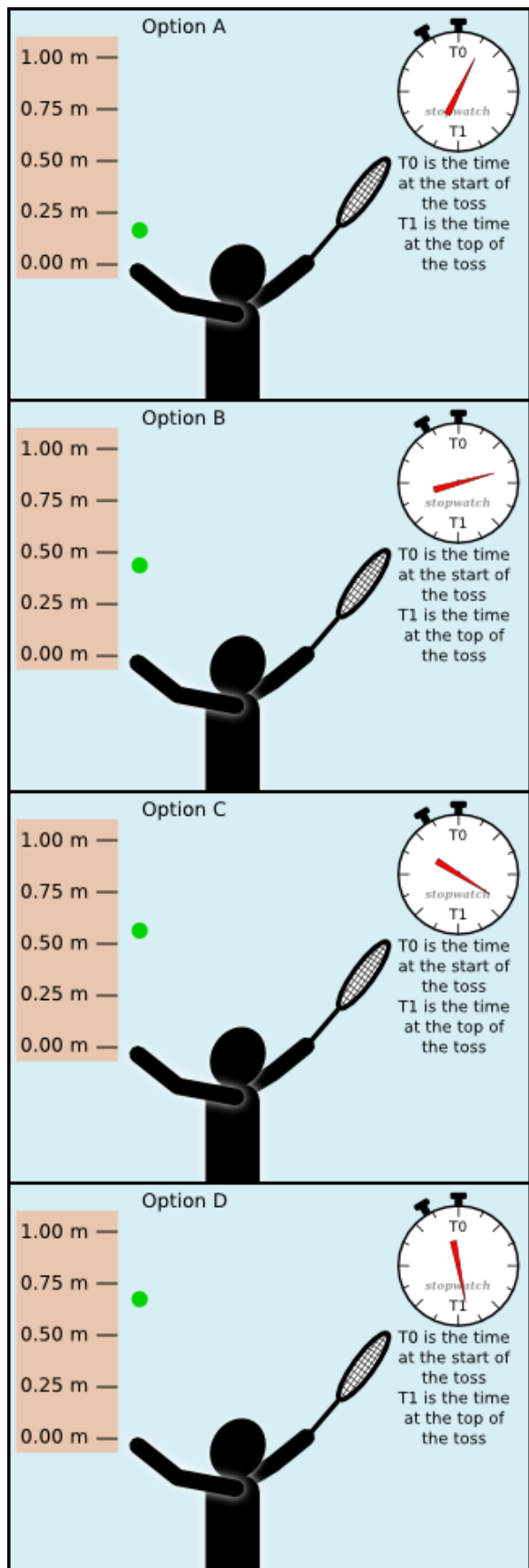
c. Assuming you have solved for time already (in part a), which of the equations below would you use to determine how fast the car was rolling as it rolled off the edge? (*KC32, KC25*)

- $v_y = v_{0y} + g * t$ (in Andes: $v_y = v0_y + g * t$)
- $v_x^2 = v_{0x}^2 + 2 * a_x * d_x$ (in Andes: $v_x^2 = v0_x^2 + 2 * a_x * d_x$)
- $d_y = v_{0y} * t + 1/2 * g * t^2$ (in Andes: $d_y = v0_y * t + 1/2 * g * t^2$)
- $d_x = v_{0x} * t$ (in Andes: $d_x = v0_x * t$)

d. When does the car begin falling? (*KC24*)

- Immediately after it leaves the table
- A few moments after it leaves the table
- It depends on how fast it was rolling
- It never falls

9. In a game of tennis, the player serving the ball tossed the tennis ball straight into the air before hitting it with her racket. Consider the trajectory of the tennis ball shown in the figure when answering the following questions. Ignore the effects of air resistance.



- a. Where is the magnitude of the vertical position largest?
 - Position A
 - Position B
 - Position C
 - Position D
 - All positions equal

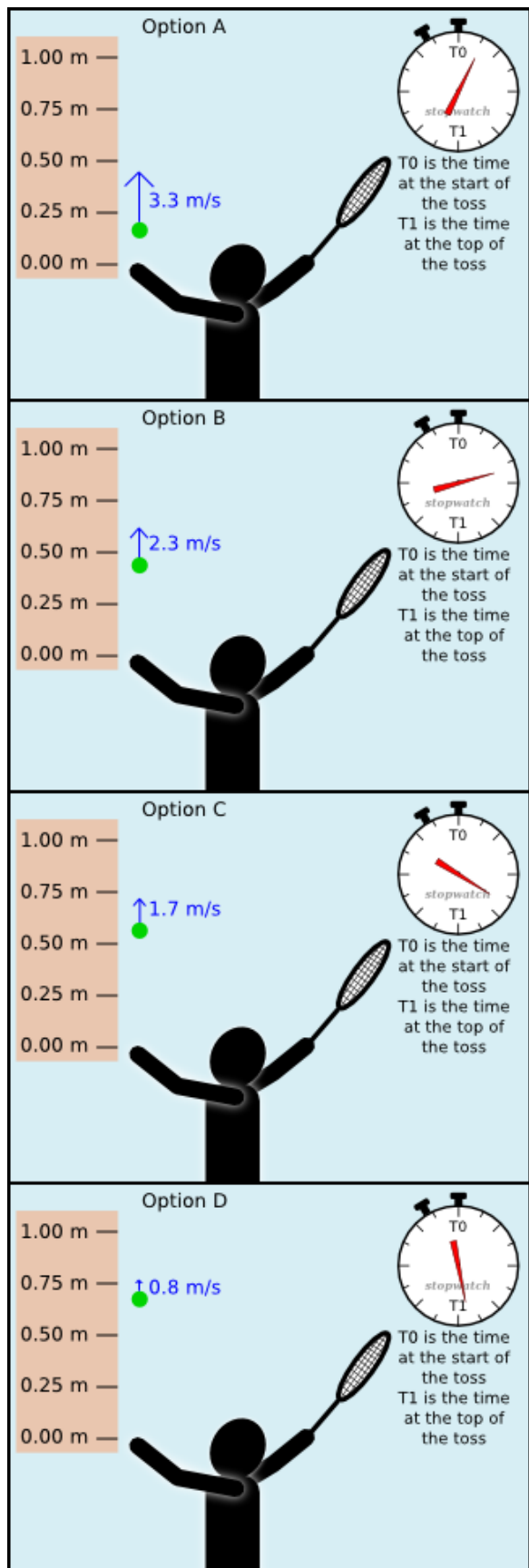
- b. Where is the magnitude of vertical position smallest?
 - Position A
 - Position B
 - Position C
 - Position D
 - All positions equal

- c. Where is the magnitude of the vertical velocity largest? (*KC28*)
 - Position A
 - Position B
 - Position C
 - Position D
 - All positions equal

- d. Where is the magnitude of vertical velocity smallest? (*KC28*)
 - Position A
 - Position B
 - Position C
 - Position D
 - All positions equal

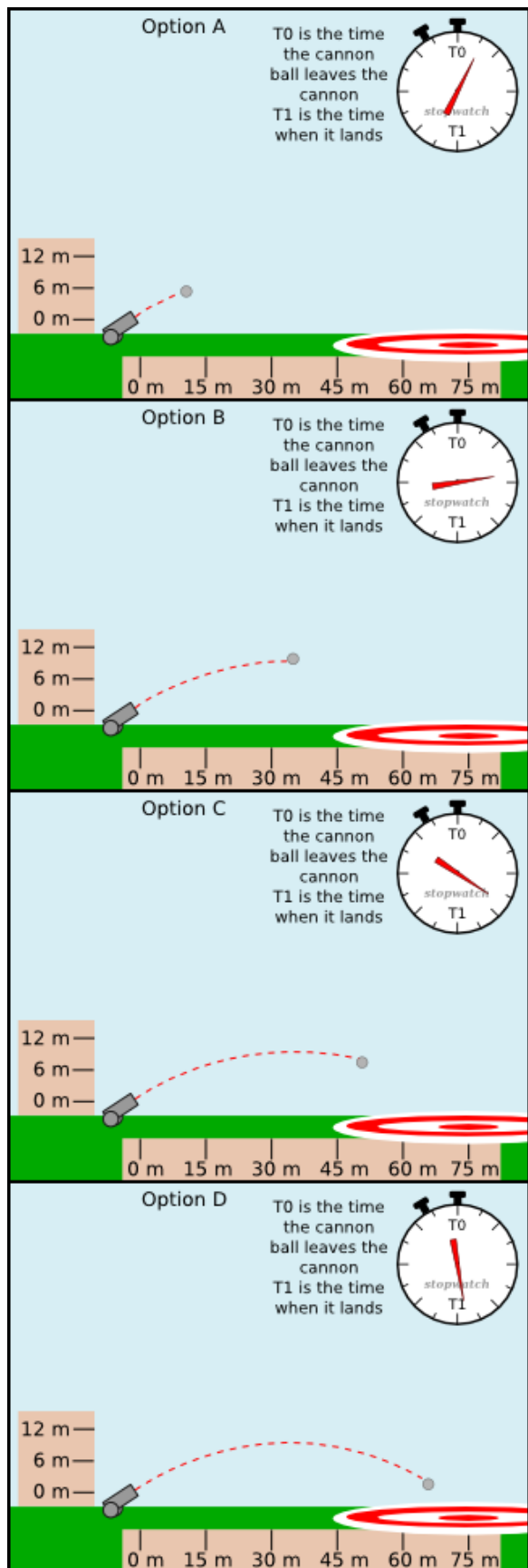
10. The server in a game of tennis tosses the ball into the (just like on the previous page). Consider these figures below when answering the following questions. Ignore the effects

of air resistance.



- a. Between time A and time B, what is the direction of vertical velocity?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- b. Between time A and time B, what is the direction of vertical acceleration? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically
- c. Between time C and time D, what is the direction of vertical velocity?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- d. Between time A and time B, what is the direction of vertical acceleration? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically

11. An army is practicing firing its cannon. They fire the cannon at a target that's 75 m away. Consider the trajectory of the cannon ball shown in the figures when answering the following questions. Ignore the effects of air resistance.



a. Where is the magnitude of the vertical velocity largest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

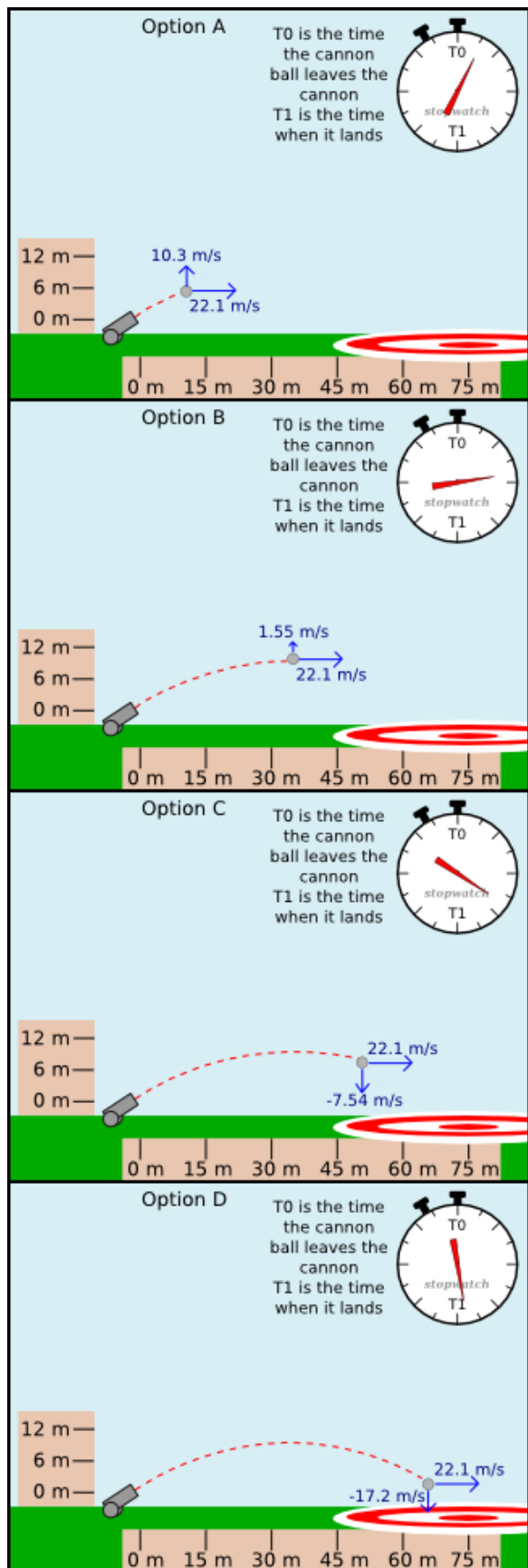
b. Where is the magnitude of the horizontal velocity largest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

c. Where is the magnitude of vertical velocity smallest? (*KC28*)

- Position A
- Position B
- Position C
- Position D
- All positions are equal

12. An army is practicing firing its cannon by firing it at a target that's 75 m away. (just like on the previous page). Consider these figures below when answering the following questions. Ignore the effects of air resistance.



- a. Between time A and time B, what is the direction of vertical **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- b. Between time A and time B, what is the direction of vertical **acceleration**? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically
- c. Between time C and time D, what is the direction of vertical **velocity**?
- up
 - down
 - left
 - right
 - it isn't moving vertically
- d. Between time C and time D, what is the direction of vertical **acceleration**? (*KC15, KC23*)
- up
 - down
 - left
 - right
 - it isn't moving vertically

e. Between time A and time B, what is the direction of horizontal **velocity**?

- up
- down
- left
- right
- it isn't moving horizontally

f. Between time A and time B, what is the direction of horizontal **acceleration**?

(*KC15*, *KC23*)

- up
- down
- left
- right
- it isn't moving horizontally

APPENDIX H

TUTORING PROBLEMS FOR THE SECOND STUDY

This appendix shows the four tutoring problems and their reflection questions for the second study, described in Chapter 6.

H.1 PROBLEM 1: SUPERTANKER

Problem statement: A modern supertanker is gigantic: 1200 to 1300 ft long with a 200 ft beam. Fully loaded, it chugs along at about 16.0 knots (8.33 m/s). It can take 1200 s to bring it to a full stop. Calculate the corresponding acceleration in m/s^2 and determine the displacement of the tanker as it slows to a stop.

Reflection Questions:

1. Is the acceleration pointing in the same direction or in the opposite direction of the supertanker's motion?
2. If the acceleration in this problem were doubled, how would the displacement change? Would it be more, less, or the same?

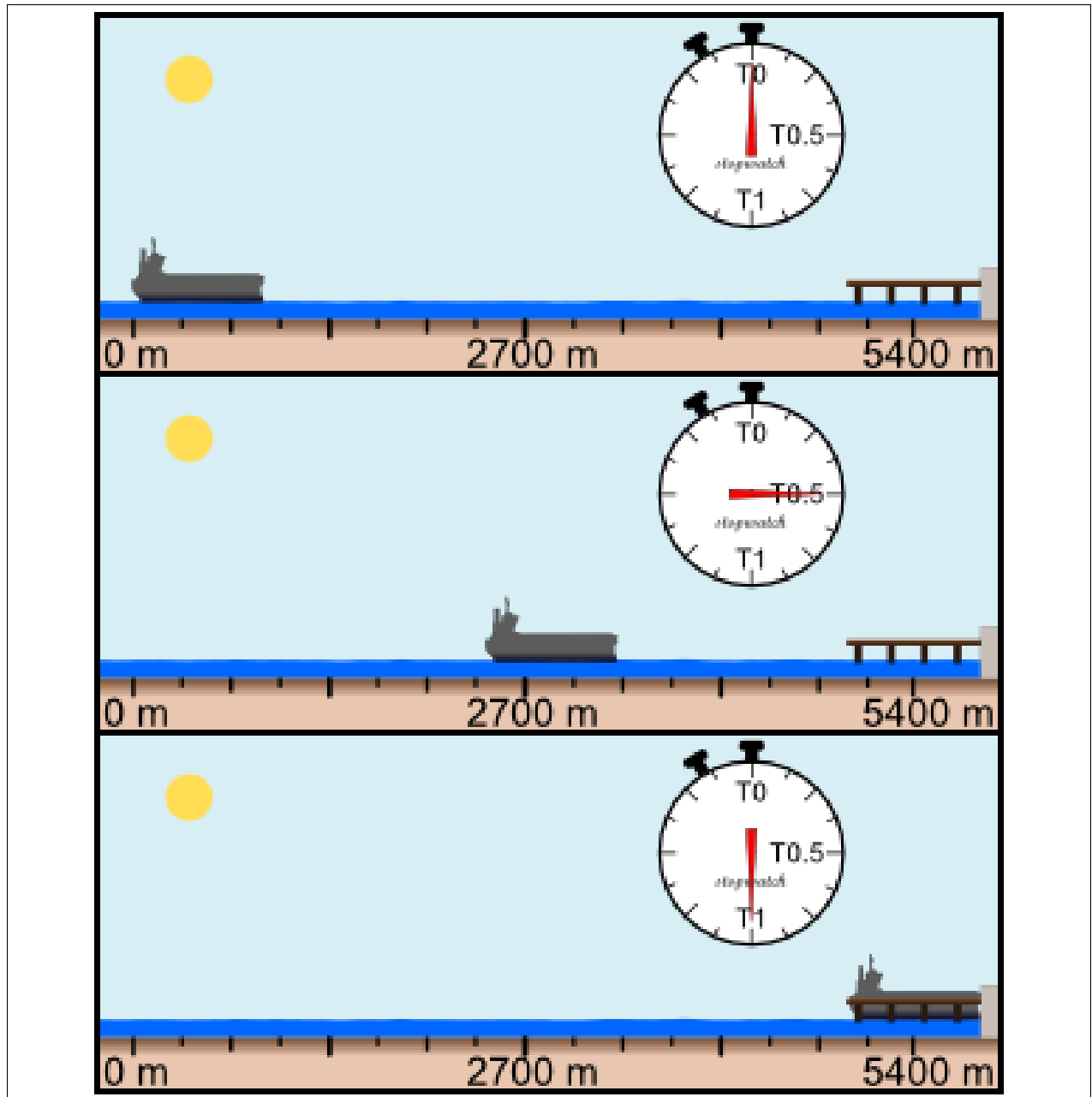


Figure H1: The illustration presented with the supertanker problem statement.

H.2 PROBLEM 2: KANGAROO

Problem statement: A kangaroo can jump about 1.50 m straight up. What is the magnitude of the take-off velocity?

Reflection Questions:

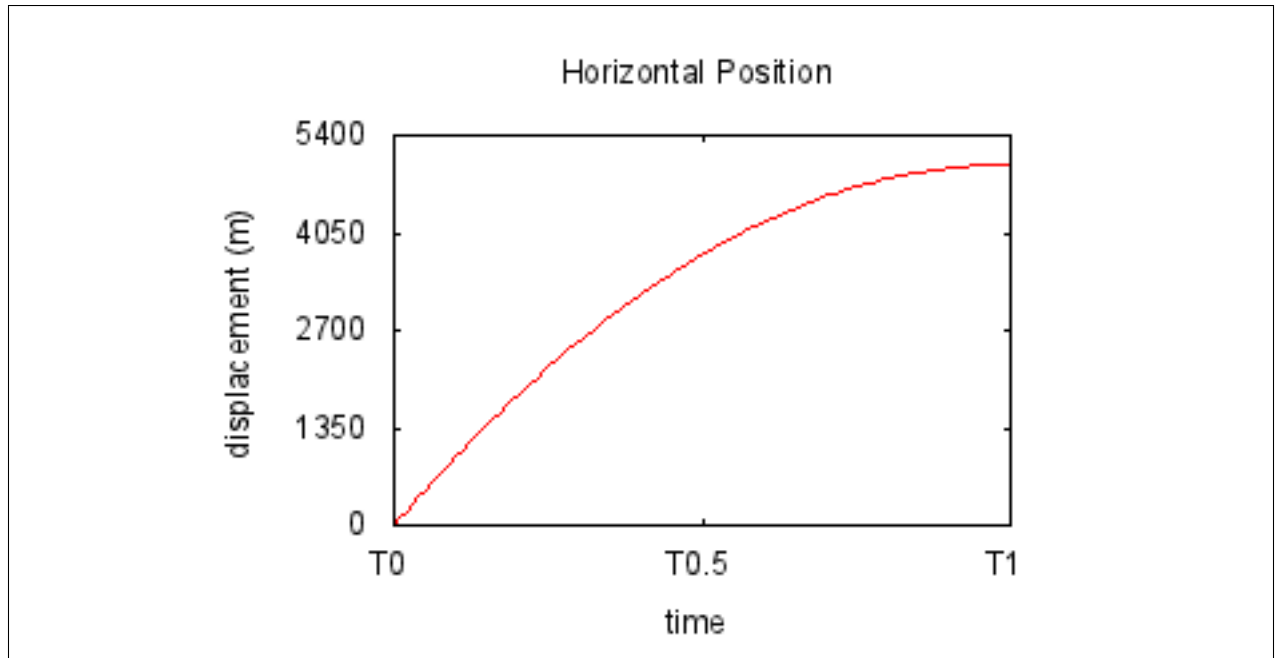


Figure H2: The graph presented with the supertanker problem statement.

1. At what time during the jump (from the moment it leaves the ground to the moment it reaches the top of the jump) is the kangaroo moving fastest?
2. What is the direction of acceleration
 - a. while the kangaroo is going up?
 - b. while the kangaroo is going down?
 - c. at the peak of the jump?

H.3 PROBLEM 3: ARROW

Problem statement: Suppose you aim a bow horizontally, directly at the center of a target 25.0 m away from you. If the speed of the arrow is 60 m/s, how far from the center of the target will it strike the target? That is, find the vertical displacement of the arrow while it is in flight. Assume there is no air friction.

Reflection Questions:

1. Instead of hitting the center of the target, the arrow was too far above or below? Why?

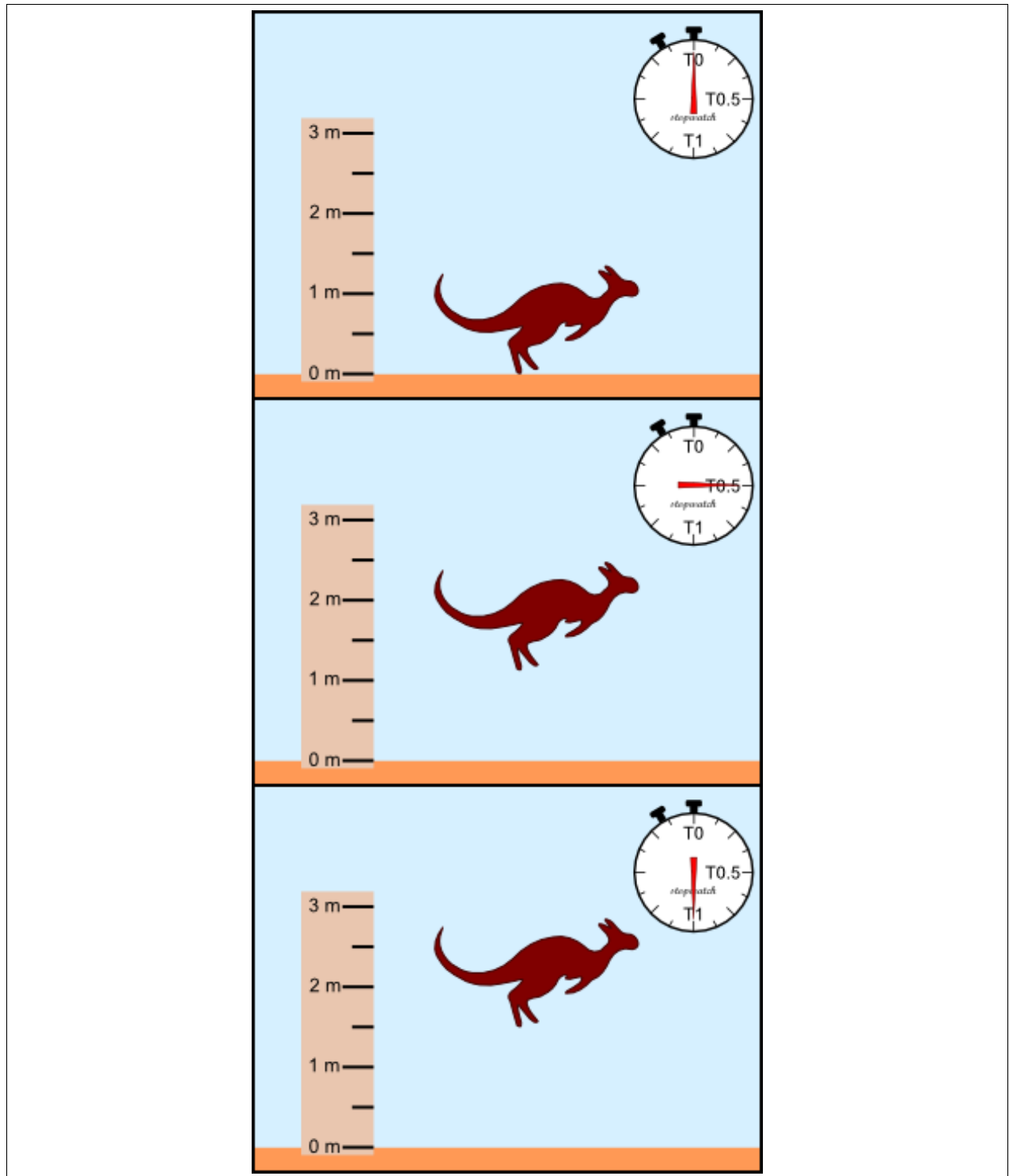


Figure H3: The illustration presented with the kangaroo problem statement.

2. What variable can the archer change so that he would hit the target?
3. Did the arrow experience an acceleration in the x-direction after being released from the

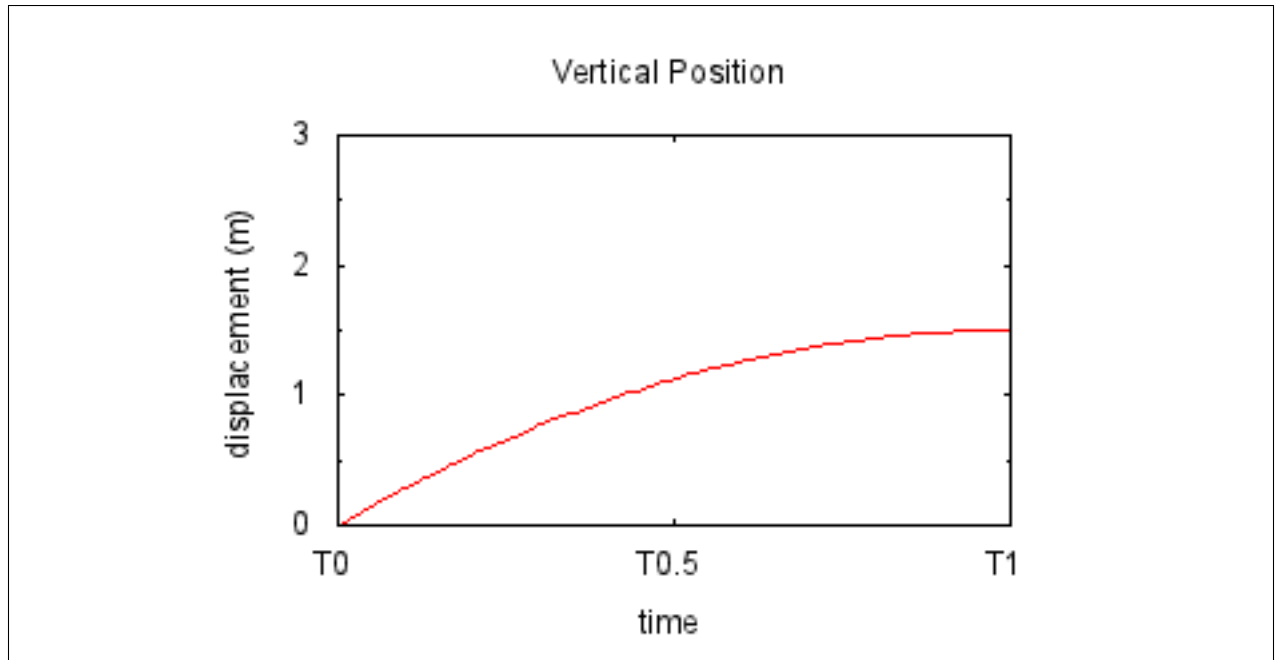


Figure H4: The graph presented with the kangaroo problem statement.

bow? Why not?

H.4 PROBLEM 4: STONE

Problem statement: A red colored stone is thrown horizontally at a velocity of 5.0 m/s from the roof of a 35.0 m building and later hits the ground below. What is the red stone's horizontal displacement? Ignore the effects of air friction.

Reflection Questions:

1. For this discussion, we will consider a new situation where the person threw the stone with a larger initial horizontal velocity. If the stone is thrown with a larger horizontal initial velocity, what happens to the time spent in the air?
2. In the original problem given, when will the direction of the stone's fall be straight down?

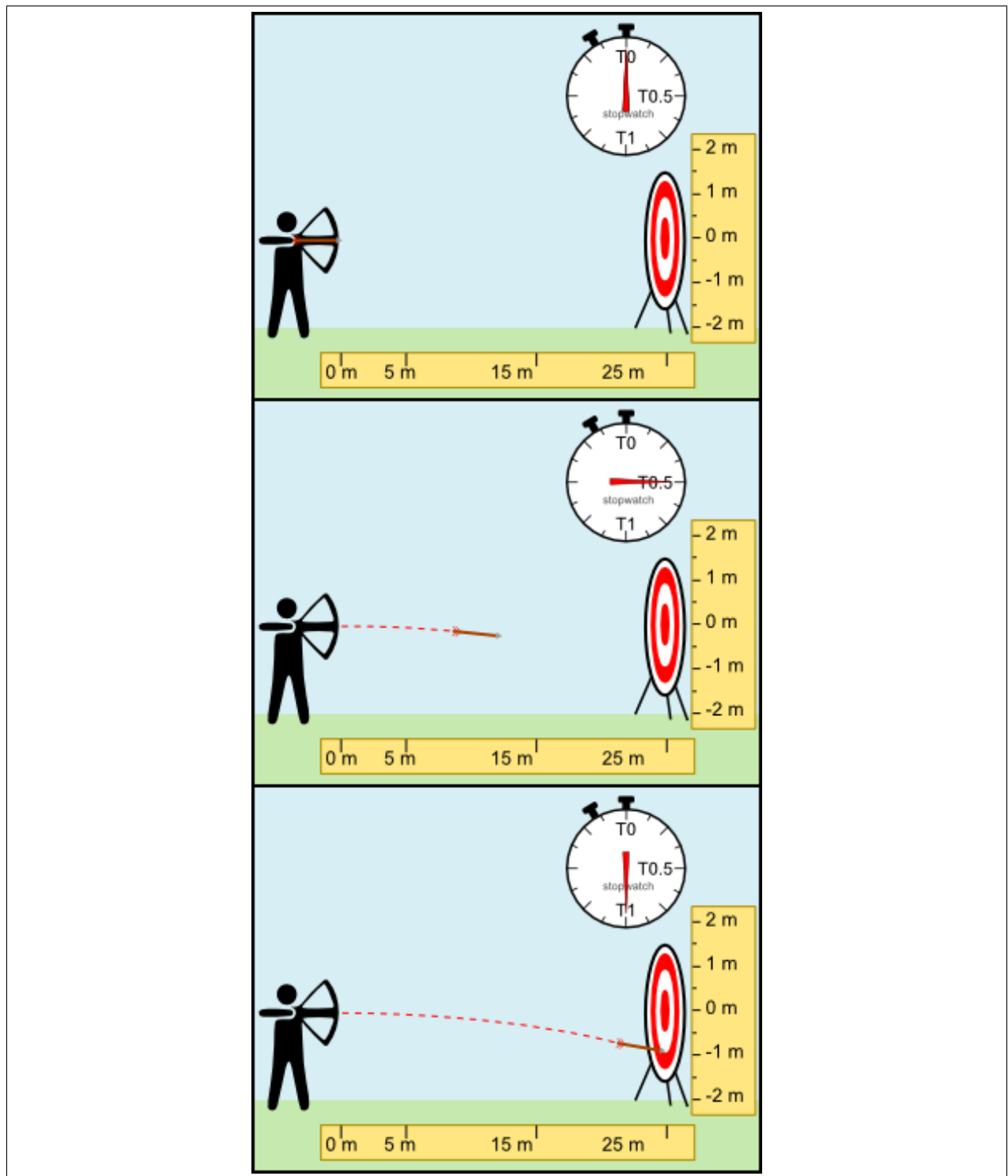


Figure H5: The illustration presented with the arrow problem statement.

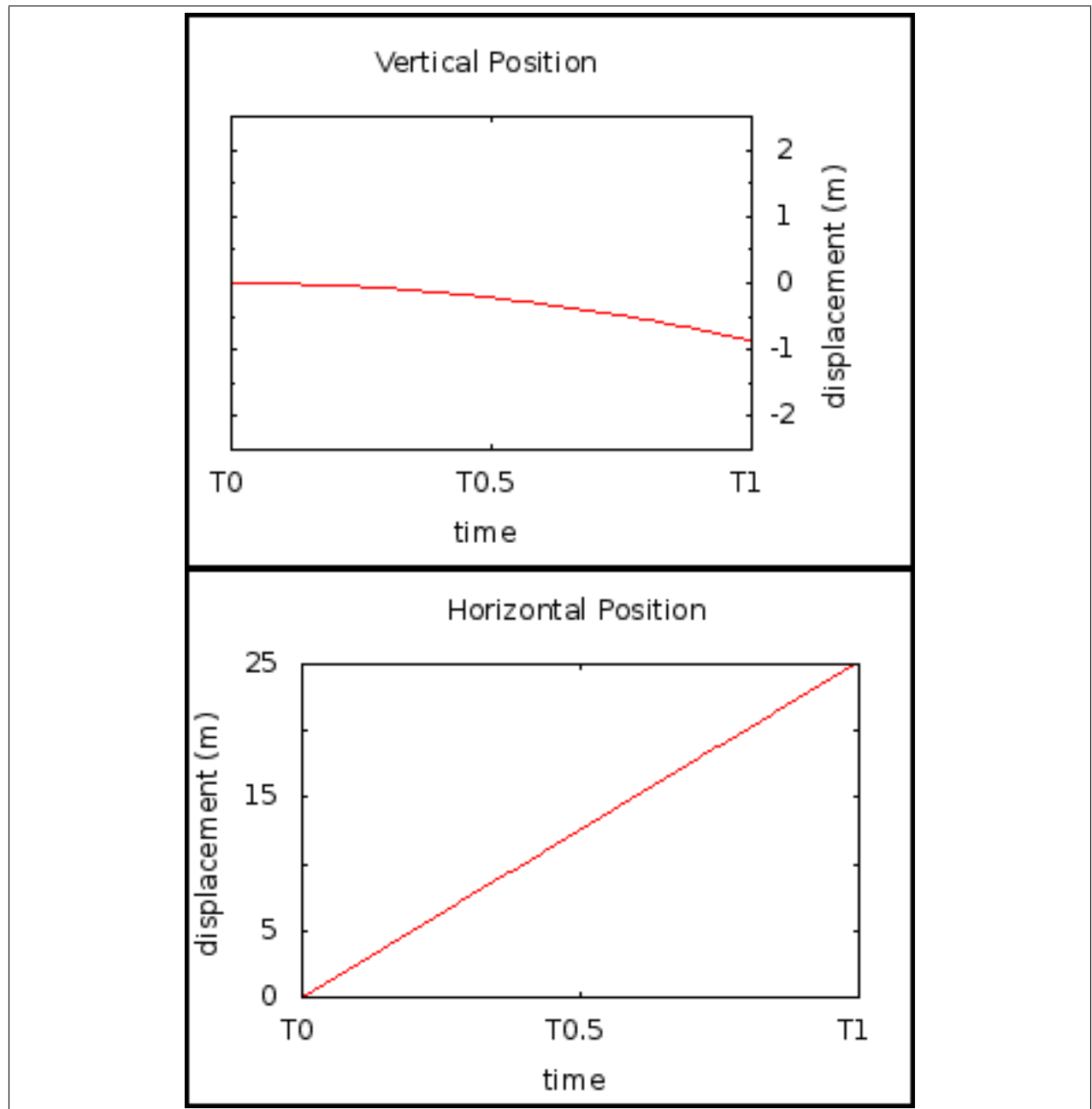


Figure H6: The graph presented with the arrow problem statement.

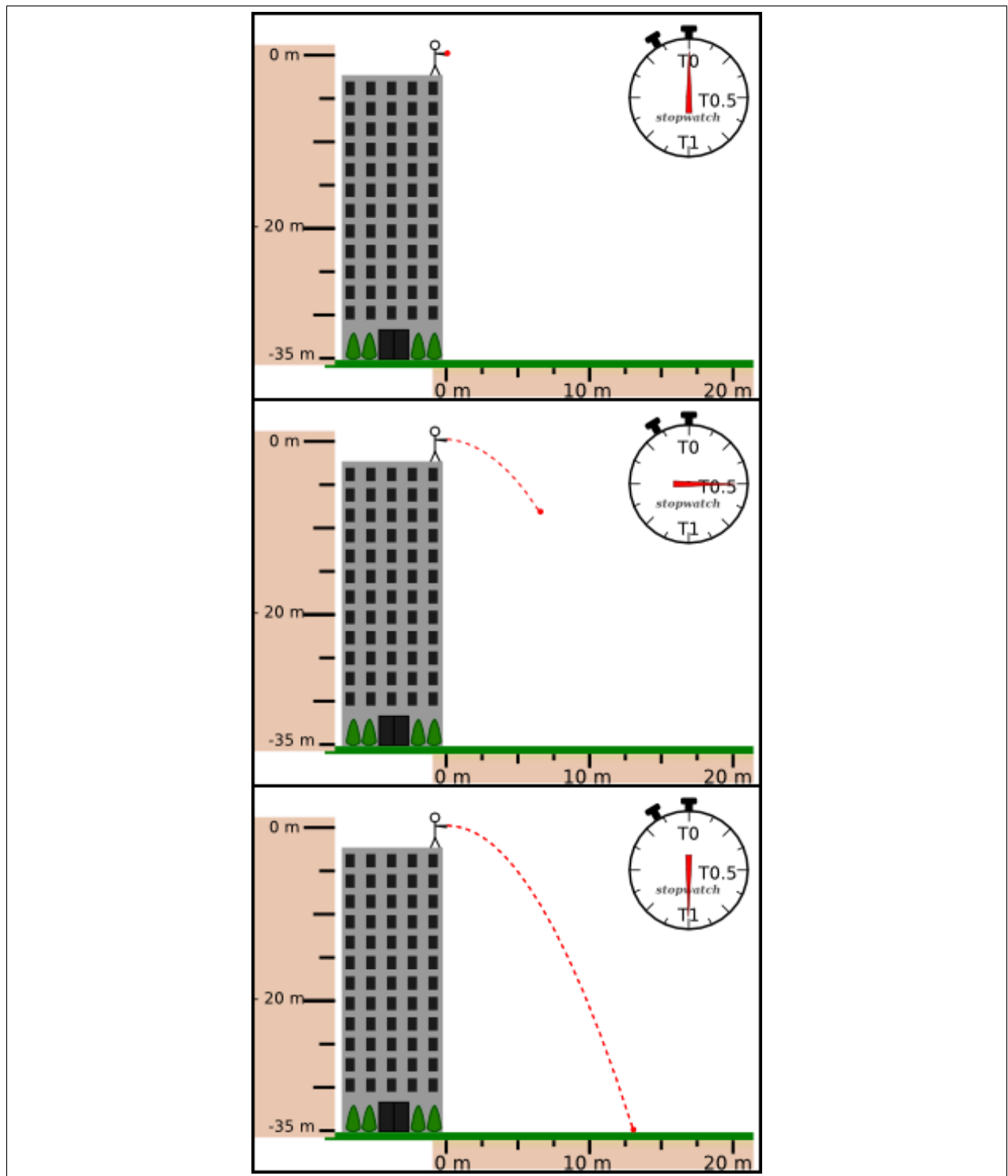


Figure H7: The illustration presented with the stone problem statement.

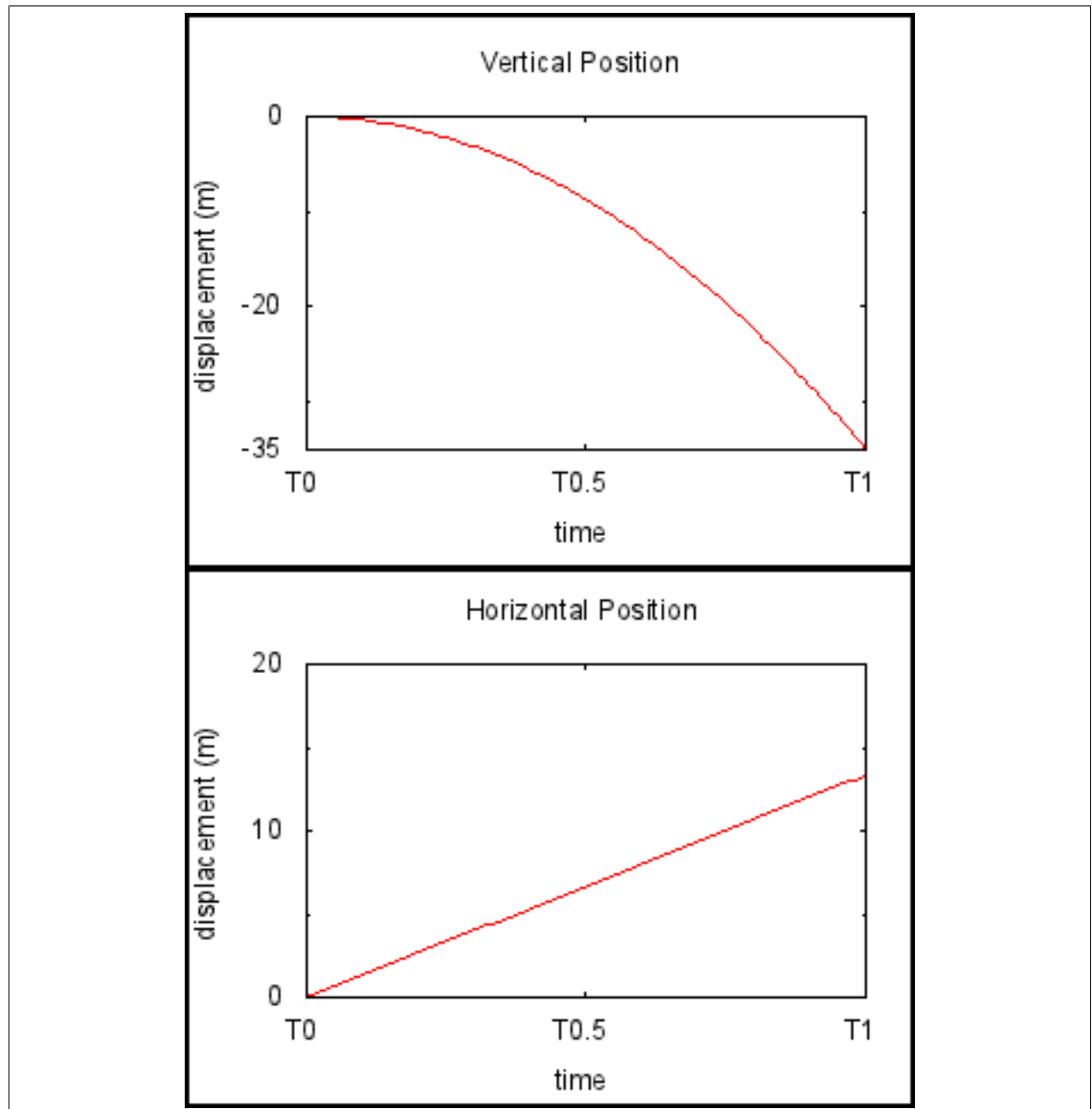


Figure H8: The graph presented with the stone problem statement.

BIBLIOGRAPHY

- [1] Obama pushes math, science education, CNN, 2009.
- [2] United states. public law 107-110. 107th congress, 2002.
- [3] W.K. Adams, S. Reid, R. LeMaster, S.B. McKagan, K.K. Perkins, M. Dubson, and C.E. Wieman. A study of educational simulations part i-engagement and learning. *Journal of Interactive Learning Research*, 19(3):397–419, 2008a.
- [4] W.K. Adams, S. Reid, R. LeMaster, S.B. McKagan, K.K. Perkins, M. Dubson, and C.E. Wieman. A study of educational simulations part ii-interface design. *Journal of Interactive Learning Research*, 19(4):551–577, 2008b.
- [5] S. Ainsworth. Deft: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3):183–198, 2006.
- [6] G. Aist, B. Kort, R. Reilly, J. Mostow, and R. Picard. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proceedings of Multimodal Interfaces*, pages 483 – 490, 2002.
- [7] P.L. Albacete and K. VanLehn. Evaluation the effectiveness of a cognitive tutor for fundamental physics concepts. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, pages 25–30. Mahwah, NJ: Erlbaum, 2000.
- [8] V.A. Aleven and K.R. Koedinger. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26(2):147–179, 2002.
- [9] Vincent Aleven, Amy Ogan, Octav Popescu, Cristen Torrey, and Kenneth Koedinger. Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In *Intelligent tutoring systems*, pages 443–454. Springer, 2004.
- [10] J.R. Anderson, A.T. Corbett, K.R. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207, 1995.

- [11] I. Arroyo, J. Beck, B. Woolf, C. Beal, and K. Schultz. Macroadaptating animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. In *Intelligent Tutoring Systems*, pages 574–583. Springer, 2000.
- [12] Ivon Arroyo, Beverly Park Woolf, Winslow Burelson, Kasia Muldner, Dovon Rai, and Minghui Tai. A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 24(4):387–426, 2014.
- [13] R. Baker. Data mining for education. *International Encyclopedia of Education*, 7:112–118, 2010.
- [14] R. Baker, A. Corbett, K. Koedinger, S. Evenson, I. Roll, A. Wagner, M. Naim, J. Raspat, D. Baker, and J. Beck. Adapting to when students game an intelligent tutoring system. In *Intelligent Tutoring Systems*, pages 392–401. Springer, 2006.
- [15] R.S. Baker, A.T. Corbett, and K.R. Koedinger. Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems*, pages 54–76. Springer, 2004.
- [16] R.S. Baker, A.T. Corbett, K.R. Koedinger, and A.Z. Wagner. Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 383–390. ACM, 2004.
- [17] G.M. Baschera, A. Busetto, S. Klingler, J. Buhmann, and M. Gross. Modeling engagement dynamics in spelling learning. In *Artificial Intelligence in Education*, pages 31–38. Springer, 2011.
- [18] J.E. Beck. Using response times to model student disengagement. In *Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments*, pages 13–20, 2004.
- [19] R.J. Beichner. Testing student interpretation of kinematics graphs. *Am. J. Phys*, 62:8, 1994.
- [20] Olesya Blazhenkova and Maria Kozhevnikov. Visual-object ability: A new dimension of non-verbal intelligence. *Cognition*, 117(3):276 – 301, 2010.
- [21] B.S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984.
- [22] H. Cen, K. Koedinger, and B. Junker. Learning factors analysis—a general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems*, pages 164–175. Springer, 2006.
- [23] C. Chen. Individual differences in a spatial-semantic virtual environment. *Journal of the American Society for Information Science*, 51(6):529–542, 2000.

- [24] M. Chi, K. Koedinger, G. Gordon, P. Jordan, and K. VanLehn. Instructional factors analysis: A cognitive model for multiple instructional interventions, 2011.
- [25] M. Chi, K. VanLehn, and D. Litman. Do micro-level tutorial decisions matter: applying reinforcement learning to induce pedagogical tutorial tactics. In *Intelligent Tutoring Systems*, pages 224–234. Springer, 2010.
- [26] M. Chi, K. VanLehn, D. Litman, and P. Jordan. An evaluation of pedagogical tutorial tactics for a natural language tutoring system: a reinforcement learning approach. *International Journal of Artificial Intelligence in Education*, 21(1):83–113, 2011.
- [27] Min Chi. Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. Master’s thesis, University of Pittsburgh, 2009.
- [28] Min Chi, Pamela Jordan, Kurt Vanlehn, and Diane Litman. To elicit or to tell: Does it matter? *Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, 200:197, 2009.
- [29] M.T.H. Chi, N. De Leeuw, M.H. Chiu, and C. LaVancher. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477, 1994.
- [30] Peter A. Cohen, James A. Kulik, and Chen-Lin C. Kulik. Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19(2):pp. 237–248, 1982.
- [31] C. Conati and H. Maclaren. Exploring the role of individual differences in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 199–206. ACM, 2008.
- [32] C. Conati and K. Vanlehn. Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education (IJAIED)*, 11:389–415, 2000.
- [33] Cristina Conati and Heather Maclaren. Modeling user affect from causes and effects. In *User Modeling, Adaptation, and Personalization*, pages 4–15. Springer, 2009.
- [34] Cristina Conati, Kasia Muldner, and Giuseppe Carenini. From example studying to problem solving via tailored computer-based meta-cognitive scaffolding: hypotheses and design. *Journal of Technology, Instruction, Cognition & Learning - Special Issue on Problem Solving Support in Intelligent Tutoring Systems*, 4(2), 2006.
- [35] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.
- [36] A.T. Corbett and J.R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

- [37] Mark G Core, Johanna D Moore, and Claus Zinn. Supporting constructive learning with a feedback planner. In *the Proceedings of the AAAI 2000 Fall Symposium, Building Dialog Systems for Tutorial Applications. Technical Report FS-00-01. AAAI Press. Menlo Park, CA. Cummins, DD, Kintsch, W., Reusser, K. & Weimer*, pages 405–438, 2000.
- [38] Ryan SJ d Baker, Albert T Corbett, and Vincent Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems*, pages 406–415. Springer, 2008.
- [39] Melissa H. Dancy and Robert Beichner. Impact of animation on assessment of conceptual understanding in physics. *Phys. Rev. ST Phys. Educ. Res.*, 2:010104, Mar 2006.
- [40] M. Dennis, J. Masthoff, and C. Mellish. Adapting performance feedback to a learners conscientiousness. *User Modeling, Adaptation, and Personalization*, pages 297–302, 2012.
- [41] A. Dillon. Spatial-semantics: How users derive shape from information space. *Journal of the American Society for Information Science*, 51(6):521–528, 2000.
- [42] S. D’Mello and A. Graesser. Affect detection from human-computer dialogue with an intelligent tutoring system. In *Intelligent Virtual Agents*, pages 54–67. Springer, 2006.
- [43] Sidney K D’Mello and Arthur C Graesser. Malleability of students’ perceptions of an affect-sensitive tutor and its influence on learning. In *FLAIRS Conference*, 2012.
- [44] S.K. D’Mello and A. Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20:147–187, 2010.
- [45] J. Drummond and D. Litman. In the zone: Towards detecting student zoning out using supervised machine learning. In *Intelligent Tutoring Systems*, pages 306–308. Springer, 2010.
- [46] Myroslava O Dzikovska, Johanna D Moore, Natalie Steinhauser, and Gwendolyn Campbell. Exploring user satisfaction in a tutorial dialogue system. In *Proceedings of the SIGDIAL 2011 Conference*, pages 162–172. Association for Computational Linguistics, 2011.
- [47] R.B. Ekstrom, J.W. French, H.H. Harman, and D. Dermen. *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service, 1976.
- [48] Martha W Evens, Ru-Charn Chang, Yoon Hee Lee, Leem Seop Shim, Chong Woo Woo, Yuemei Zhang, Joel A Michael, and Allen A Rovick. Cirsim-tutor: An intelligent tutoring system using natural language dialogue. In *Proceedings of the fifth conference*

- on *Applied natural language processing: Descriptions of system demonstrations and videos*, pages 13–14. Association for Computational Linguistics, 1997.
- [49] K. Forbes-Riley and D. Litman. Adapting to student uncertainty improves tutoring dialogues. In *Proceeding of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 33–40, 2009.
 - [50] K. Forbes-Riley and D. Litman. Adapting to multiple affective states in spoken dialogue. *Proceedings of the 13th Annual Meeting of the Special Interest Group on on Discourse and Dialogue (SIGDIAL)*, pages 217–226, 2012.
 - [51] K. Forbes-Riley, D. Litman, H. Friedberg, and J. Drummond. Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In *Proc. NAACL-HLT*, 2012.
 - [52] Kate Forbes-Riley and Diane Litman. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language*, 25(1):105 – 126, 2011. jce:titleAffective Speech in Real-Life Interactionsi/ce:titlej.
 - [53] Kate Forbes-Riley and Diane Litman. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language*, 25(1):105–126, 2011.
 - [54] Kate Forbes-Riley and Diane Litman. When does disengagement correlate with performance in spoken dialog computer tutoring? *International Journal of Artificial Intelligence in Education*, 22(1):39–58, 2013.
 - [55] Douglas C. Giancoli. *Physics: Principles with Applications (7th Edition)*. Addison-Wesley, 2013.
 - [56] J. Gobert, M. Sao Pedro, R. Baker, E. Toto, and O. Montalvo. Leveraging educational data mining for real time performance assessment of scientific inquiry skills within micro-worlds. *Journal of Educational Data Mining*, 2012.
 - [57] R.L. Goldstone and J.Y. Son. The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences*, 14(1):69–110, 2005.
 - [58] A.C. Graesser, P. Chipman, B.C. Haynes, and A. Olney. Autotutor: an intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on*, 48(4):612 – 618, nov. 2005.
 - [59] A.C. Graesser, S. Lu, G.T. Jackson, H.H. Mitchell, M. Ventura, A. Olney, and M.M. Louwerse. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods*, 36(2):180–192, 2004.
 - [60] I.A. Halloun and D. Hestenes. The initial knowledge state of college physics students. *American journal of Physics*, 53(11):1043–1055, 1985.

- [61] N. Heffernan and K. Koedinger. An intelligent tutoring system incorporating a model of an experienced human tutor. In *Intelligent Tutoring Systems*, pages 596–608. Springer, 2002.
- [62] T. Hirashima, I. Imai, T. Horiguchi, and T. Toumoto. Error-based simulation to promote awareness of errors in elementary mechanics and its evaluation. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 409–416. IOS Press, 2009.
- [63] G.T. Jackson, A. Olney, A.C. Graesser, and HJ Kim. Autotutor 3-d simulations: Analyzing users actions and learning trends. In *Proceedings of the 28th annual meeting of the Cognitive Science Society*, pages 1557–1562, 2006.
- [64] P. Jordan, B. Hall, M. Ringenberg, Y. Cui, and C.P. Rosé. Tools for authoring a dialogue agent that participates in learning studies. pages 43–50, 2007.
- [65] S. Katz, P. Jordan, D. Litman, and the Rimac Project Team. Rimac: A natural-language dialogue system that engages students in deep reasoning. 2011.
- [66] Sandra Katz, David Allbritton, and John Connelly. Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education*, 13(1):79–116, 2003.
- [67] Kenneth R Koedinger and Vincent Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3):239–264, 2007.
- [68] K.R. Koedinger, J.R. Anderson, W.H. Hadley, and M.A. Mark. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education (IJAIED)*, 8:30–43, 1997.
- [69] Patrick B. Kohl and Noah D. Finkelstein. Student representational competence and self-assessment when solving physics problems. *Phys. Rev. ST Phys. Educ. Res.*, 1:010104, Oct 2005.
- [70] M. Kozhevnikov, M.A. Motes, and M. Hegarty. Spatial visualization in physics problem solving. *Cognitive Science*, 31(4):549–579, 2007.
- [71] C.L.C. Kulik, J.A. Kulik, and R.L. Bangert-Drowns. Effectiveness of mastery learning programs: A meta-analysis. *Review of educational research*, 60(2):265–299, 1990.
- [72] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, Sheelagh Carpendale, et al. Seven guiding scenarios for information visualization evaluation. Technical report, Department of Computer Science, University of Calgary, 2011.
- [73] J.J. LaViola Jr and R.C. Zeleznik. Mathpad 2: a system for the creation and exploration of mathematical sketches. In *ACM SIGGRAPH 2007 courses*, page 46. ACM, 2007.

- [74] K. Leelawong and G. Biswas. Designing learning by teaching agents: The betty’s brain system. *International Journal of Artificial Intelligence in Education*, 18(3):181–208, 2008.
- [75] M. Lipschultz and D. Litman. Illustrations or Graphs: Some Students Benefit From One Over the Other. In *Artificial Intelligence in Education*, 2013.
- [76] M. Lipschultz and D. Litman. Modeling Student Benefit from Illustrations and Graphs. In *International Conference on Intelligent Tutoring Systems*, pages 436–441, 2014.
- [77] M. Lipschultz, D. Litman, P. Jordan, and S. Katz. Predicting changes in level of abstraction in tutor responses to students. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-2011)*, 2011.
- [78] D.J. Litman, C.P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembé, and S. Silliman. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, 16(2):145–170, 2006.
- [79] Adrian Maries, Nathaniel Mays, Megan J. Olson Hunt, Kim F. Wong, William Layton, Caterina Rosano, Howard Aizenstein, Robert Boudreau, S. Levent Yilmaz, and G. Elisabeta Marai. Grace: A visual comparison framework for integrated spatial and non-spatial geriatric data. *IEEE TRANSACTIONS ON COMPUTER GRAPHICS AND VISUALIZATION*, 6(1):274–285, (under review).
- [80] Brent Martin, Antonija Mitrovic, Kenneth R Koedinger, and Santosh Mathan. Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction*, 21(3):249–283, 2011.
- [81] L.C. McDermott, M.L. Rosenquist, and E.H. vanZee. Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, 55(6):503–513, 1987.
- [82] D.S. McNamara and W. Kintsch. Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes*, 22(3):247–288, 1996.
- [83] Nicole M. McNeil and Emily R. Fyfe. “concreteness fading” promotes transfer of mathematical knowledge. *Learning and Instruction*, 22(6):440 – 448, 2012.
- [84] D.E. Meltzer. Relation between students problem-solving performance and representational format. *American Journal of Physics*, 73:463, 2005.
- [85] Allen Newell and Paul S Rosenbloom. Mechanisms of skill acquisition and the law of practice. In Anderson J., editor, *Cognitive skills and their acquisition*, Lecture Notes in Computer Science. Erlbaum, 1981.
- [86] P.I. Pavlik, H. Cen, and K.R. Koedinger. Performance factors analysis—a new alternative to knowledge tracing. In *Proceedings of the 2009 conference on Artificial*

- Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 531–538. IOS Press, 2009.
- [87] H. Pon-Barry, K. Schultz, E.O. Bratt, B. Clark, and S. Peters. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, 16(2):171–194, 2006.
 - [88] Stephen Provasnik, Patrick Gonzales, and David Miller. Us performance across international assessments of student achievement: Special supplement to the condition of education 2009. nces 2009-083. *National Center for Education Statistics*, 2009.
 - [89] M. Rau, V. Aleven, and N. Rummel. Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions. In *Intelligent Tutoring Systems*, pages 413–422. Springer, 2010.
 - [90] M. Rau, V. Aleven, N. Rummel, and S. Rohrbach. Sense making alone doesnt do it: Fluency matters too! its support for robust learning with multiple representations. In *Intelligent Tutoring Systems*, pages 174–184. Springer, 2012.
 - [91] M. Rau, N. Rummel, V. Aleven, L. Pacilio, and Z. Tunc-Pekkan. How to schedule multiple graphical representations? a classroom experiment with an intelligent tutoring system for fractions. In *The future of learning: Proceedings of the 10th international conference of the learning sciences (ICLS 2012)*, pages 64–71, 2012.
 - [92] M.A. Rau, V. Aleven, and N. Rummel. Intelligent tutoring systems with multiple representations and self-explanation prompts support learning of fractions. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 441–448. IOS Press, 2009.
 - [93] A. Renkl. Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21(1):1–29, 1997.
 - [94] Jeff Rickel, Neal Lesh, Charles Rich, C Sidner, and Abigail Gertner. Building a bridge between intelligent tutoring and collaborative dialogue systems. In *Proceedings of Tenth International Conference on Artificial Intelligence in Education*, pages 592–594, 2001.
 - [95] C.P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein. Interactive conceptual tutoring in atlas-andes. In *Proceedings of AI in Education 2001 Conference*, pages 151–153, 2001.
 - [96] K. Scheiter, P. Gerjets, and J. Schuh. The acquisition of problem-solving skills in mathematics: How animations can aid understanding of structural problem features and solution procedures. *Instructional Science*, 38(5):487–502, 2010.
 - [97] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

- [98] B. Shih, K.R. Koedinger, and R. Scheines. A response time model for bottom-out hints as worked examples. *Educational Data Mining*, page 117, 2008.
- [99] C. Singh. Effect of misconception on transfer in problem solving. In *AIP Conference Proceedings*, volume 951, page 196, 2007.
- [100] C. Singh. Coupling conceptual and quantitative problems to develop expertise in introductory physics students. In *AIP Conference Proceedings*, volume 1064, page 199, 2008.
- [101] Dana Tenneson. Chempad: A pedagogical tool for exploring handwritten organic molecules. Master’s thesis, Brown University, 2005.
- [102] D. Toker, C. Conati, G. Carenini, and M. Haraty. Towards adaptive information visualization: on the influence of user characteristics. *User Modeling, Adaptation, and Personalization*, pages 274–285, 2012.
- [103] A. Van Heuvelen. Learning to think like a physicist: A review of research-based instructional strategies. *Am. J. Phys.*, 59:891–897, 1991.
- [104] A. Van Heuvelen and X. Zou. Multiple representations of work–energy processes. *American Journal of Physics*, 69:184–194, 2001.
- [105] K. VanLehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.
- [106] K. VanLehn, A.C. Graesser, G.T. Jackson, P. Jordan, A. Olney, and C.P. Rosé. When are tutorial dialogues more effective than reading? *Cognitive Science*, 31(1):3–62, 2007.
- [107] K. VanLehn, P. Jordan, and D. Litman. Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In *Proceedings of SLaTE Workshop on Speech and Language Technology in Education ISCA Tutorial and Research Workshop*, pages 17–20, 2007.
- [108] K. VanLehn, P. Jordan, C. Rosé, D. Bhembé, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, et al. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Intelligent tutoring systems*, pages 158–167. Springer, 2002.
- [109] K. VanLehn, C. Lynch, K. Schulze, J.A. Shapiro, R. Shelby, D. Treacy, and M. Wintersgill. The andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, 15:147–204, 2005.
- [110] M.C. Velez, D. Silver, and M. Tremaine. Understanding visualization through spatial ability differences. In *Visualization*, pages 511–518. IEEE, 2005.

- [111] J. Walonoski and N. Heffernan. Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In *Intelligent Tutoring Systems*, pages 382–391. Springer, 2006.
- [112] A. Ward, J. Connelly, S. Katz, D. Litman, and C. Wilson. Cohesion, semantics and learning in reflective dialog. In *AIED 2009: 14 th International Conference on Artificial Intelligence in Education Workshops Proceedings*, page 18, 2010.
- [113] A. Ward and D. Litman. Cohesion and learning in a tutorial spoken dialog system. In *Proceedings of the 19th International FLAIRS Conference (FLAIRS-19)*, pages 533–538, 2006.
- [114] A. Ward and D. Litman. Semantic cohesion and learning. In *Intelligent Tutoring Systems*, pages 459–469. Springer, 2008.
- [115] Beverly Woolf, Toby Dragon, Ivon Arroyo, David Cooper, Winslow Burleson, and Kasia Muldner. Recognizing and responding to student affect. In JulieA. Jacko, editor, *Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction*, volume 5612 of *Lecture Notes in Computer Science*, pages 713–722. Springer Berlin Heidelberg, 2009.
- [116] Beverly Woolf, Matt Romoser, Dan Bergeron, and Don Fisher. Tutoring 3-dimensional visual skills: Dynamic adaptation to cognitive level. In *International Conference on Artificial Intelligence and Education*, 2003.
- [117] B.P. Woolf. A roadmap for education technology. 2010.
- [118] C. Ziemkiewicz, R.J. Crouser, A.R. Yauilla, S.L. Su, W. Ribarsky, and R. Chang. How locus of control influences compatibility with visualization style. In *Visual Analytics Science and Technology (VAST)*, pages 81–90. IEEE, 2011.
- [119] C. Ziemkiewicz and R. Kosara. Preconceptions and individual differences in understanding visual metaphors. In *Computer Graphics Forum*, volume 28, pages 911–918. Wiley Online Library, 2009.
- [120] Paul W. Zitzewitz, Todd George Elliott, David G. Haase, Kathleen A. Harper, Michael R. Herzog, Jane Bray Nelson, Jim Nelson, Charles A. Schuler, and Margaret K. Zorn. *Physics Principles and Problems*. McGraw Hill Glencose, Columbus, Ohio, 2005.